Likelihood-Free Hypothesis Testing and Applications of the Energy Distance

by

Patrik Róbert Gerber

MMath, University of Oxford (2019)

Submitted to the Department of Mathematics
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN MATHEMATICS AND STATISTICS

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2024

Authored by:	Patrik Róbert Gerber Department of Mathematics May 3, 2024
Certified by:	Philippe Rigollet Professor of Mathematics, Thesis Supervisor
Accepted by:	Jonathan Kelner Professor of Mathematics Graduate Chair, Applied Mathematics

Likelihood-Free Hypothesis Testing and Applications of the Energy Distance

by

Patrik Róbert Gerber

Submitted to the Department of Mathematics
on May 3, 2024 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN MATHEMATICS AND STATISTICS

ABSTRACT

This thesis studies questions in nonparametric testing and estimation that are inspired by machine learning. One of the main problems of our interest is likelihood-free hypothesis testing: given three samples X, Y and Z with sample sizes n, n and m respectively, one must decide whether the distribution of Z is closer to that of X or that of Y. We fully characterize the problem's sample complexity for multiple distribution classes and with high probability. We uncover connections to two-sample, goodness-of-fit and robust testing, and show the existence of a trade-off of the form $mn \asymp k/\epsilon^4$, where k is an appropriate notion of complexity and ϵ is the total variation separation between the distributions of X and Y. We generalize our problem to allow Z to come from a mixture of the distributions of X and Y, and propose a kernel-based test for its solution, and also verify the existence of a trade-off between m and n on experimental data from particle physics. In addition, we demonstrate that the family of "classifier accuracy" tests are not only popular in practice but also provably near-optimal, recovering and simplifying a multitude of classical and recent results. Finally, we study affine classifiers as a tool for estimation and testing, with the key technical tool being a connection to the energy distance. In particular, we propose a density estimation routine based on minimizing the generalized energy distance, targeting smooth densities and Gaussian mixtures. We interpret our results in terms of half-space separability over these classes, and derive analogous results for discrete distributions. As a consequence we deduce that any two discrete distributions are well-separated by a half-space, provided their support is embedded as a packing of a high-dimensional unit ball. We also scrutinize two recent applications of the energy distance in the two-sample testing literature.

Thesis supervisor: Philippe Rigollet
Title: Professor of Mathematics

Acknowledgments

I would like to thank my academic mentors Prof. Philippe Rigollet and Prof. Yury Polyanskiy. I thank Philippe for welcoming me into his research group, for instilling in me his desire to pursue good problems and for believing in me. I thank Yury for his enthusiasm, kindness and for helping me find my footing.

The great advantage of studying at MIT was the opportunity to learn from the brightest and most dedicated people I have ever met. During the past five years I had the fortune of collaborating with Jason Altschuler, Sinho Chewi, Thibaut Le Gouic, Yanjun Han, Tianze Jiang, Holden Lee, Chen Lu, Austin Stromme, Rui Sun and Paxton Turner. Without fail, I have learnt valuable lessons from each of them, mathematical and otherwise, which will serve me wherever life may take me, and for which I am grateful.

In addition to the friends I have made through my research, I must also thank my friends outside of work who were there with me during both the good times and the bad times. I will miss our nightly discussions with Calder and Matthew in 71 5^{th}, and all the jokes we have shared. I will miss the early morning outings with Adam in his street-legal car, and I will miss our arguments with David. I will also miss the many other friends I've made over the years, including Cameron, Catherine, Deeparaj, Elisabetta, Felipe, George, Heather, Khashayar, Marisa, Mitchell, Sujit and Tina. I am thankful to my friends back home in Kecskemét for always welcoming me after my travels. My frequent calls commiserating with Máté were essential in surviving the height of the pandemic.

I am thankful for the unconditional love and support of my family: Christoph, Brigi, Mom, Dad, Mama, Papa and Grosi. They have shaped me into who I am today, and they share in all of my successes. I am eternally grateful to my fiancée Panka for always giving me something to look forward to. I love you and I am excited about what our future holds.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The present chapter serves as an overview of the results of this thesis which, broadly speaking, studies questions in nonparametric testing and estimation that are inspired by machine learning. The material presented is based on the papers [77, 76, 74, 75].

1.1 Structure of the Thesis

The first paper [77] resolves the minimax sample complexity of "likelihood-free hypothesis testing" (LFHT), which is a simplified model of a problem that emerges in the growing field of likelihood-free inference (LFI). The results are summarized in Section 1.4, and the paper can be found in full detail in Chapter 2.

The second paper [76] studies a generalization of LFHT inspired by particle physics experiments, and uses an approach based on kernels. We also perform experiments with trained kernels parametrized by neural networks, supporting our theoretical results. This is summarized in Section 1.6, and the full paper can be found in Chapter 3.

The third paper [74] studies classifier-accuracy testing, which is essentially a meta-algorithm that is widely used by practitioners of LFI. The results are summarized in Section 1.5 and Sections 1.5.4 and 1.5.5 in particular, and the paper is reproduced in Chapter 4 in full.

The fourth paper [75] studies the use of affine classifiers for nonparametric testing and estimation, with the key technical tool being a connection to the energy distance. These results are also summarized in Section 1.5, and Sections 1.5.1 to 1.5.3 in particular. The paper these sections are based on can be found in Chapter 5.

The remaining sections of this chapter are as follows. Section 1.2 presents the technical preliminaries on nonparametric testing, estimation, LFHT and the energy distance that are needed to state our results. Finally, Section 1.7 gives a short summary of our approach to proving minimax lower bounds for testing problems in general, and LFHT in particular.

1.2 Technical Preliminaries

Here we go over the technical preliminaries required to state and understand our results. In Section 1.2.1 we formally introduce the problem of likelihood-free hypothesis testing (LFHT)

which is the main object of study in much of this thesis. In Section 1.2.2 we introduce some classical statistical problems that LFHT is related to. Section 1.2.3 introduces the distribution classes that are studied throughout Sections 1.4 to 1.7. Finally, Section 1.2.4 serves as an introduction to the generalized energy distance, which is further studied in Section 1.5.

1.2.1 Definition of LFHT

Suppose we observe three i.i.d. samples $X_1, \dots, X_n$, $Y_1, \dots, Y_n$ and $Z_1, \dots, Z_m$ from P_X, P_Y and P_Z respectively, taking values in a common set $\mathcal{X}$. Suppose in addition that $\mathcal{P}$ is a set of probability distributions on $\mathcal{X}$ and that $\epsilon \in (0,1)$. Likelihood-free hypothesis testing (LFHT) over the class $\mathcal{P}$ with separation ϵ is defined as the problem of testing the hypothesis

$$\begin{aligned} H_0(\epsilon, \mathcal{P}) &= \Big\{(P_X, P_Y, P_Z) \in \mathcal{P}^3 : \mathsf{TV}(P_X, P_Y) \geq \epsilon, P_X = P_Z\Big\} \\ &\qquad\qquad \text{versus} \qquad\qquad (\mathsf{LFHT}) \\ H_1(\epsilon, \mathcal{P}) &= \Big\{(P_X, P_Y, P_Z) \in \mathcal{P}^3 : \mathsf{TV}(P_X, P_Y) \geq \epsilon, P_Y = P_Z\Big\}. \end{aligned}$$

Here we regard a hypothesis as a set of probability measures from which the data may have come. Given $\delta \in (0, 1/2)$, we say that the "test", that is, the measurable function $\Psi_{n,m,\epsilon,\delta} : \mathcal{X}^{2n+m} \to \{0,1\}$ performs LFHT with worst-case error δ over the class $\mathcal{P}$ with separation ϵ, if

$$\max_{i=0,1} \sup_{(P_X,P_Y,P_Z)\in H_i(\epsilon,\mathcal{P})} \mathbb{P}\big(\Psi_{n,m,\epsilon,\delta}(X_1, \dots, X_n, Y_1, \dots, Y_n, Z_1, \dots, Z_m) \neq i\big) \leq \delta. \tag{1.2.1}$$

We also allow the test $\Psi_{n,m,\epsilon,\delta}$ to be a non-deterministic function of the data, that is, to take an independent random seed as input, however we suppress this in the notation. Given the parameters $\epsilon, \delta, \mathcal{P}$, we can define the region

$$\mathcal{R}_{\mathsf{LF}}(\epsilon, \delta, \mathcal{P}) := \Big\{(n,m) \in \mathbb{N}^2 : \text{there exists a test } \Psi_{n,m,\epsilon,\delta} \text{ satisfying (1.2.1)}\Big\}, \tag{1.2.2}$$

which represents the set of sample sizes (n, m) for which it is possible to perform LFHT at maximum error δ even in the worst-case over the distributions P_X, P_Y, or using other words, in a minimax sense. It is of primary interest to us to characterize the region $\mathcal{R}_{\mathsf{LF}}(\epsilon, \delta, \mathcal{P})$ up to constants independent of ϵ, δ. As we shall see, in the case of discrete distributions we also study the dependence on the support size, which will be denoted k, in addition to ϵ and δ.

Remark 1. *Our definition of LFHT above was inspired by likelihood-free inference (LFI), as we shall explain in Section 1.4.1 further. In the context of LFI, one may think of the X_i and Y_i as the data generated by our simulator, and the Z_i as the experimental data. A priori we don't have a good model of the likelihood of our data (P_Z) or the output distribution of the simulation (P_X and P_Y); we only know that they lie in some class $\mathcal{P}$ and are separated in total variation.*

Remark 2. *LFHT has been proposed and studied before our work under other names, see Section 1.3.2 for historical details.*

Remark 3. *When studying the constant error a.k.a. $\delta = \Theta(1)$ regime, we will often simply write $\delta = 0.3$. Note that the value 0.3 could be replaced by any non-trivial error probability, that is, any value strictly less than 1/2. This is because a simple strategy that splits data into $\log(1/\delta)$ disjoint and equal sized batches and then takes a majority vote of the output of an optimal constant error test achieves worst-case error $\mathcal{O}(\delta)$ at the cost of a multiplicative $\log(1/\delta)$ factor in the sample sizes n, m. We shall see later that the δ-dependence produced by this naive strategy is sub-optimal.*

1.2.2 Four Fundamental Problems in Statistics

As in the previous section, assume that $\mathcal{P}$ is a family of probability distributions throughout.

Binary Hypothesis Testing

First we consider binary hypothesis testing, one of the fundamental problems in statistics. Given i.i.d. observed data $Z_1, \dots, Z_m$ with law P_Z and candidate distributions P_0, P_1, the task is to decide, using full knowledge of P_0, P_1, between the simple hypotheses

$$H_0 = \{P_0\} \qquad \text{versus} \qquad H_1 = \{P_1\}.$$

This problem is famously and optimally solved by the Neyman-Pearson likelihood-ratio test [157], and its error probability satisfies

$$\mathrm{err}(P_0, P_1) := \inf_{\Psi} \max_{i=0,1} \sup_{P_Z \in H_i} \mathbb{P}(\Psi(Z_1, \dots, Z_m) \neq i) = \exp(-\Theta(m\mathsf{H}^2(P_0, P_1))),$$

where H denotes the Hellinger distance, and the implied constant is universal. For a proof of this fact see [168, Section 14.6]. From the above it follows that

$$\begin{aligned} n_{\mathsf{HT}}(\epsilon, \delta, \mathcal{P}) &:= \min\left\{m \in \mathbb{N} : \sup_{P_0, P_1 \in \mathcal{P}: \mathsf{TV}(P_0, P_1) \geq \epsilon} \mathrm{err}(P_0, P_1) \leq \delta\right\} \\ &= \Theta\left(\frac{\log(1/\delta)}{\inf_{P_0, P_1 \in \mathcal{P}: \mathsf{TV}(P_0, P_1) \geq \epsilon} H^2(P_0, P_1)}\right), \end{aligned}$$

where the constant is again universal. The expression $\inf_\Psi$ denotes the infimum over all tests, possibly randomized, taking values in $\{0, 1\}$. Here n_{HT} measures the worst-case sample complexity, that is, the minimum number of observations required in the worst case to perform binary hypothesis testing with error probability δ over the class $\mathcal{P}$ under the guarantee that the two hypotheses are ϵ-separated. In all the cases that we care about n_{HT} works out to be $\Theta(\log(1/\delta)/\epsilon^2)$, and thus we rarely use the notation n_{HT}. This simplification holds whenever $\mathcal{P}$ is rich enough to include pairs of distributions (P_ϵ, Q_ϵ) with $\mathsf{TV}(P_\epsilon, Q_\epsilon) = \Theta(H(P_\epsilon, Q_\epsilon)) = \Theta(\epsilon)$ for a sequence ϵ converging to 0.

Goodness-of-Fit Testing

The second problem we consider is that of goodness-of-fit testing. Here we observe i.i.d. data $X_1, \dots, X_n$ with law P_X and the task is to decide between the hypotheses

$$H_0(\epsilon, \mathcal{P}) = \{P_0\} \qquad \text{versus} \qquad H_1(\epsilon, \mathcal{P}) = \{P \in \mathcal{P} : \mathsf{TV}(P, P_0) \geq \epsilon\},$$

where the null distribution $P_0 \in \mathcal{P}$ is know to the tester. We define the sample complexity of goodness-of-fit testing as

$$n_{\mathsf{GoF}}(\epsilon, \delta, \mathcal{P}) := \min\Big\{n \in \mathbb{N} : \sup_{P_0 \in \mathcal{P}} \inf_{\Psi} \max_{i=0,1} \sup_{P_X \in H_i(\epsilon, \mathcal{P})} \mathbb{P}(\Psi(X_1, \dots, X_n) \neq i) \leq \delta\Big\}.$$

In other words, n_{GoF} denotes the minimum number of samples required by any procedure that is able to test H_0 versus H_1 in the worst case over $P_0 \in \mathcal{P}$.

Two-Sample Testing

Next we consider two-sample testing. Here we observe two i.i.d. samples $X_1, \dots, X_n$ and $Z_1, \dots, Z_m$ with distribution P_X and P_Z respectively. The goal is to decide between the hypotheses

$$H_0(\epsilon, \mathcal{P}) = \{(P, P) : P \in \mathcal{P}\} \qquad \text{versus} \qquad H_1(\epsilon, \mathcal{P}) = \{(P, Q) \in \mathcal{P}^2 : \mathsf{TV}(P, Q) \geq \epsilon\}.$$

Notice that two-sample testing is quite similar to goodness-of-fit testing, except the null distribution is no longer known to the statistician. We define the set of n, m pairs for which two-sample testing is possible as

$$\mathcal{R}_{\mathsf{TS}}(\epsilon, \delta, \mathcal{P}) := \Big\{(n, m) \in \mathbb{N}^2 : \inf_{\Psi} \max_{i=0,1} \sup_{(P_X, P_Z) \in H_i(\epsilon, \mathcal{P})} \mathbb{P}(\Psi(X_1, \dots, X_n, Z_1, \dots, Z_m) \neq i) \leq \delta\Big\}.$$

We also define the minimax sample complexity of two-sample testing as

$$n_{\mathsf{TS}}(\epsilon, \delta, \mathcal{P}) = \min\{n : (n, n) \in \mathcal{R}_{\mathsf{TS}}(\epsilon, \delta, \mathcal{P})\},$$

which measures the difficulty of the problem when we assume the two samples are of equal size. This distinction will actually be meaningful, as it is not necessarily the case that $\mathcal{R}_{\mathsf{TS}} \asymp \{(n, m) : \min\{n, m\} \geq n_{\mathsf{TS}}\}$.

Estimation

Finally, we define the problem of estimation. Given an i.i.d. sample $X_1, \dots, X_n \sim P_X$, the goal is to produce an estimator $\widehat{P}_X$ that is close to P_X in total variation either in expectation or with high probability. We define the sample complexity of estimation to be

$$n_{\mathsf{Est}}(\epsilon, \mathcal{P}) := \min\Big\{n \in \mathbb{N} : \inf_{\widehat{P}_X} \sup_{P_X \in \mathcal{P}} \mathbb{E}\big[\mathsf{TV}(\widehat{P}_X, P_X)\big] \leq \epsilon\Big\},$$

where the minimum is taken over all (possibly randomized) estimators $\widehat{P}_X$, that is, measurable functions of the observed data $X_1, \dots, X_n$ taking values in $\mathcal{P}$.

1.2.3 Distribution Classes

Here we provide a list of the four choices of $\mathcal{P}$ that we study in Sections 1.4 and 1.5. In Section 1.6 we consider a slightly different class of distributions, defined implicitly in terms of a reproducing kernel Hilbert space (RKHS).

Smooth Densities on the Hypercube

Let $\beta > 0$ and set $\underline{\beta} := \lceil \beta - 1 \rceil$ as the largest integer strictly smaller than β. Write $\mathcal{C}(\beta, d, C)$ for the set of functions $f : [0,1]^d \to \mathbb{R}$ that are $\underline{\beta}$-times differentiable and satisfy

$$\max\left(\max_{0 \leq |\alpha| \leq \underline{\beta}} \|f^{(\alpha)}\|_\infty, \sup_{x \neq y \in [0,1]^d, |\alpha| = \underline{\beta}} \frac{|f^{(\alpha)}(x) - f^{(\alpha)}(y)|}{\|x - y\|_2^{\beta - \underline{\beta}}} \right) \leq C,$$

where $|\alpha| = \sum_{i=1}^d \alpha_i$ and $f^{(\alpha)} = \partial_1^{\alpha_1} \dots \partial_d^{\alpha_d} f$ for the multiindex $\alpha \in \mathbb{N}^d$. In other words, $\mathcal{C}(\beta, d, C)$ is the set of functions whose partial derivatives up to order $\underline{\beta}$ are bounded by C and whose partial derivatives of order $\underline{\beta}$ are $(\beta - \underline{\beta})$-Hölder continuous with constant C. Finally, define $\mathcal{P}_{\mathsf{H}}(\beta, d, C)$ to be the class of distributions with Lebesgue-densities in $\mathcal{C}(\beta, d, C)$.

We will assume throughout that $C > 1$ when referring to $\mathcal{P}_{\mathsf{H}}(\beta, d, C)$. This is to ensure that there are infinitely many distributions belonging to the class, so as to avoid vacuous statements.

Gaussian Sequence Model

Given $s, C > 0$, define the Sobolev ellipsoid $\mathcal{E}(s, C)$ of smoothness s and size C as

$$\mathcal{E}(s, C) := \Big\{ \theta \in \mathbb{R}^{\mathbb{N}} : \sum_{j=1}^{\infty} j^{2s} \theta_j^2 \leq C \Big\}.$$

Given a sequence $\theta \in \mathbb{R}^{\mathbb{N}}$, let us abuse notation and write $\mathcal{N}(\theta, 1) := \bigotimes_{i=1}^{\infty} \mathcal{N}(\theta_i, 1)$, where $\mathcal{N}(\mu, \sigma^2)$ for $\mu, \sigma \in \mathbb{R}$ denotes the one-dimensional Gaussian measure with mean μ and variance σ^2. We define our second class as

$$\mathcal{P}_{\mathsf{G}}(s, C) := \Big\{ \mathcal{N}(\theta, 1) \ : \theta \in \mathcal{E}(s, C) \Big\}.$$

The motivation for the study of $\mathcal{P}_{\mathsf{G}}$ stems from regression. Consider the classical Gaussian white noise model. Here we are given an observation of the stochastic process

$$\mathrm{d}Y_t = f(t)\mathrm{d}t + \mathrm{d}W_t, \ t \in [0,1],$$

where $(W_t)_{t \geq 0}$ denotes Brownian motion and $f \in L^2[0,1]$ is unknown. Suppose now that $\{\phi_i\}_{i \geq 1}$ forms an orthonormal basis for $L^2[0,1]$, and given an observation Y define the values

$$y_i := \langle Y, \phi_i \rangle = \int_0^1 f(t)\phi_i(t)\mathrm{d}t + \int_0^1 \phi_i(t)\mathrm{d}W_t =: \theta_i + \epsilon_i.$$

Notice that $\epsilon_i \sim \mathcal{N}(0,1)$ and that $\mathbb{E}[\epsilon_i \epsilon_j] = \mathbb{1}\{i = j\}$. In other words, the sequence $\{y_i\}_{i \geq 1}$ is an observation from the distribution $\mathcal{N}(\theta, 1)$. Consider the particular case of $\phi_1 \equiv 1$ and $\phi_{2k}(x) = \sqrt{2}\cos(2\pi k x), \phi_{2k+1}(x) = \sqrt{2}\sin(2\pi k x)$ for $k \geq 1$ and assume that f satisfies periodic boundary conditions. Then θ denotes the Fourier coefficients of f and the condition that $\sum_{j=1}^{\infty} j^{2s}\theta_j^2 \leq C$ is equivalent to an upper bound on the order $(s,2)$-Sobolev norm of f, which is loosely speaking equal to the L^2-norm of the s'th derivative of f. For a proof of this see for example [198, Proposition 1.14]. To summarize, by studying the class $\mathcal{P}_{\mathsf{G}}$ we can deduce results for signal detection or regression under Gaussian white noise, where the signal or regression function has bounded Sobolev norm.

Discrete Distributions

The third class we consider is fairly self-explanatory, it is the set of all discrete distributions. Given $k \in \mathbb{N}$, let

$$\mathcal{P}_{\mathsf{D}}(k) := \Big\{\text{distributions on the finite alphabet } \{1, 2, \dots, k\}\Big\}.$$

The fourth and final family of distributions are the "regular" or "bounded" discrete distributions. For a constant $C > 1$, define

$$\mathcal{P}_{\mathsf{Db}}(k, C) := \Big\{p \in \mathcal{P}_{\mathsf{D}}(k) : \|p\|_\infty \le C/k\Big\}.$$

Such distributions are quite natural as they show up from the discretization of continuous distributions with bounded density, such as those belonging to $\mathcal{P}_{\mathsf{H}}$.

1.2.4 Introduction to the Generalized Energy Distance

The notion of energy distance was introduced by Székely in a series of lectures at the Budapest University of Technology and Economics [195], and has been independently rediscovered multiple times [62, 211, 18].[1] Given an exponent $\gamma \in (0, 2)$ and two probability measures μ, ν on $\mathbb{R}^d$ with finite γ'th moment, the generalized energy distance between them is defined as

$$\mathcal{E}_\gamma(\mu, \nu) = \sqrt{2\mathbb{E}\|X - Y\|^\gamma - \mathbb{E}\|X - X'\|^\gamma - \mathbb{E}\|Y - Y'\|^\gamma}, \tag{1.2.3}$$

where $(X, X', Y, Y') \sim \mu^{\otimes 2} \otimes \nu^{\otimes 2}$ and $\|\cdot\|$ denotes the Euclidean norm. The energy distance, and related ideas such as "distance covariance/correlation" (which likely first appeared in [71]), have been applied to many problems in statistics with success, such as goodness-of-fit [188, 174, 159, 38, 140, 150], two-sample [192, 46, 99, 171, 18] and independence testing [194, 62, 191, 194, 189], see also the book [193] for additional references. The asymptotic properties of these methods have been understood deeply, however, the energy distance hasn't enjoyed much attention in non-asymptotic statistics. Here, we will be concerned with the latter. Since the energy distance is a Maximum Mean Discrepancy (MMD) [180, Theorem 22], generic results about MMDs can be specialized to the energy distance. Nevertheless, by studying it directly we can derive previously unnoticed properties that lead to efficient estimators and tests, and discover interesting connections to half-spaces along the way. First, let us give a short exposition of some important properties of $\mathcal{E}_\alpha$, that will be used later on.

Representation as a Sobolev Norm

Given a probability measure μ on $\mathbb{R}^d$, define its Fourier transform as

$$\widehat{\mu}(\omega) = \int_{\mathbb{R}^d} e^{-i\langle x, \omega\rangle} \mathrm{d}\mu(x) = \mathbb{E}_{X\sim\mu}\big[e^{-i\langle \omega, X\rangle}\big].$$

We have the following simple analytic fact: given any $\omega \in \mathbb{R}^d$ and $\gamma \in (0, 2)$,

$$\int_{\mathbb{R}^d} \frac{1 - \cos\left(\langle x, \omega\rangle\right)}{\|x\|^{d+\gamma}} \mathrm{d}x = \|\omega\|^\gamma \frac{\pi^{d/2}\Gamma(1 - \gamma/2)}{\gamma 2^{\gamma-1}\Gamma(\frac{d+\gamma}{2})}. \tag{1.2.4}$$

[1] Despite nontrivial effort I haven't been able to obtain a copy of [195].

Indeed, using a rotation we may write the integral as

$$\int_{\mathbb{R}^d} \frac{1-\cos\langle x,\omega\rangle}{\|x\|^{d+\gamma}}\mathrm{d}x = \int_{\mathbb{R}^d} \frac{1-\cos\left(x_1\|\omega\|\right)}{\|x\|^{d+\gamma}}\mathrm{d}x = \|\omega\|^\gamma \int_{\mathbb{R}^d} \frac{1-\cos(x_1)}{\|x\|^{d+\gamma}}\mathrm{d}x.$$

The above integral is clearly finite: since $1-\cos(t) \sim t^2/2$ at the origin, the integrand behaves like $1/\|x\|^{d-\gamma} \ll 1/\|x\|^d$ at zero, and its tail behaves like $1/\|x\|^{d+\gamma} \ll 1/\|x\|^d$ as $x \to \infty$. The conclusion follows by an explicit computation which we omit here, but details can be found for example in [190, Appendix A]. Using Equation (1.2.4) we can show that $\mathcal{E}_\gamma$ is a weighted L^2-distance between the Fourier transforms $\widehat{\mu}$ and $\widehat{\nu}$. Alternatively, one may also recognize the formula as the homogenous Sobolev norm of order $-(d+\gamma)/2$, up to constant.

Proposition 1.2.1 ([190, Proposition 2])**.** *Let $\gamma \in (0,2)$ and μ, ν be probability measures with finite γ'th moment.*

$$\mathcal{E}_\gamma^2(\mu,\nu) = \frac{1}{C(d,\gamma)} \int_{\mathbb{R}^d} \frac{\left|\widehat{\mu}(\omega) - \widehat{\nu}(\omega)\right|^2}{\|\omega\|^{d+\gamma}}\mathrm{d}\omega,$$

where $C(d,\gamma) = \frac{\pi^{d/2}\Gamma(1-\gamma/2)}{\gamma 2^{\gamma-1}\Gamma(\frac{d+\gamma}{2})}$.

Proof. It is easier if we work backwards from the final result. Following the proof from [190], and writing $(X, X', Y, Y') \sim \mu^{\otimes 2} \otimes \nu^{\otimes 2}$, we have

$$\begin{aligned}
&\int_{\mathbb{R}^d} \frac{|\widehat{\mu}(\omega) - \widehat{\nu}(\omega)|^2}{\|\omega\|^{d+\gamma}}\mathrm{d}\omega \qquad (1.2.5)\\
&\quad= \int_{\mathbb{R}^d} \mathbb{E}\left[\frac{(e^{i\langle X,\omega\rangle} - e^{i\langle Y,\omega\rangle})(e^{-i\langle X',\omega\rangle} - e^{-i\langle Y',\omega\rangle})}{\|\omega\|^{d+\gamma}}\right]\mathrm{d}\omega\\
&\quad= \int_{\mathbb{R}^d} \mathbb{E}\left[\frac{\cos\langle X-X',\omega\rangle - \cos\langle Y-X',\omega\rangle - \cos\langle X-Y',\omega\rangle + \cos\langle Y-Y',\omega\rangle}{\|\omega\|^{d+\gamma}}\right]\mathrm{d}\omega\\
&\quad\overset{\text{Eq. (1.2.4)}}{=} C(d,\gamma)\mathcal{E}_\gamma^2(\mu,\nu),
\end{aligned}$$

where the second to last line used that $\sin(x) = -\sin(-x)$. □

Representation as a Maximum Mean Discrepancy

The second interpretation of the energy distance is as a Maximum Mean Discrepancy (MMD). First, let us give a non-rigorous crash course on the background required to understand our results. For a complete treatment we refer the readers to the book [96].

Given a set $\mathcal{X}$, we call the function $K : \mathcal{X}^2 \to \mathbb{R}$ a kernel if the $n \times n$ matrix with ij'th entry $K(x_i, x_j)$ is symmetric positive semidefinite for all choices of $x_1, \ldots, x_n \in \mathcal{X}$ and $n \geq 1$. There is a unique reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ associated to K. $\mathcal{H}_K$ consists of functions $\mathcal{X} \mapsto \mathbb{R}$ and satisfies the reproducing property $\langle K(x,\cdot), f\rangle_{\mathcal{H}_K} = f(x)$ for all $f \in \mathcal{H}_k$ and $x \in \mathcal{X}$, in particular $K(x,\cdot) \in \mathcal{H}_K$. Given a probability measure P on $\mathcal{X}$, define its kernel embedding θ_P as

$$\theta_P := \mathbb{E}_{X\sim P} K(X,\cdot) = \int_{\mathcal{X}} K(x,\cdot)P(\mathrm{d}x). \qquad (1.2.6)$$

Given the kernel embeddings of two probability measures P, Q, we can measure their distance in the RKHS by $\mathrm{MMD}(P,Q) := \|\theta_P - \theta_Q\|_{\mathcal{H}_K}$, where MMD stands for maximum mean discrepancy. MMD has a closed form thanks to the reproducing property and linearity:

$$\mathrm{MMD}^2(P,Q) = \mathbb{E}\Big[K(X,X') + K(Y,Y') - 2K(X,Y)\Big],$$

where $(X,X',Y,Y') \sim P^{\otimes 2} \otimes Q^{\otimes 2}$. In particular, if P, Q are empirical measures based on observations, we can evaluate the MMD exactly in quadratic time, which is crucial in practice. Yet another attractive property of MMD is that (under mild integrability conditions) it is an integral probability metric (IPM) where the supremum is over the unit ball of the RKHS $\mathcal{H}_K$, that is,

$$\mathrm{MMD}(P,Q) = \sup_{f \in \mathcal{H}_K : \|f\|_{\mathcal{H}_K} \le 1} \int f(x)(\mathrm{d}P(x) - \mathrm{d}Q(x)),$$

see for example [179, 151].

In the case of the generalized energy distance, the kernel is given by

$$k_\gamma(x,y) = \|x\|^\gamma + \|y\|^\gamma - \|x-y\|^\gamma, \tag{1.2.7}$$

which in one dimension corresponds to the covariance operator of fractional Brownian motion. For a proof of the fact that k_γ above is positive definite see [180, Proposition 3]. With k_γ defined as above, it is trivial to observe that the generalized energy distance $\mathcal{E}_\gamma$ is equal to the MMD with kernel k_γ. The following result is a simple consequence of the MMD formulation of $\mathcal{E}_\gamma$.

Lemma 1.2.2. *Let ν be a probability distribution on $\mathbb{R}^d$ and let $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ for an i.i.d. sample $X_1, \dots, X_n$ from ν. Then, for any $\gamma \in (0,2)$,*

$$\mathbb{E}\mathcal{E}_\gamma^2(\nu, \nu_n) \le \frac{2M_\gamma(\nu)}{n}.$$

Proof. Let $\tilde{X}_1, \dots, \tilde{X}_n$ be an additional i.i.d. sample from ν, and write $\tilde{\nu}_n$ for the corresponding empirical measure. Using the definition of $\mathcal{E}_\gamma$ in (1.2.3), we can compute

$$\begin{aligned}\mathbb{E}\mathcal{E}_\gamma^2(\nu_n, \tilde{\nu}_n) = \mathbb{E}\Big[\frac{2}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|X_i - \tilde{X}_j\|^\gamma - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|\tilde{X}_i - \tilde{X}_j\|^\gamma \\ - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|X_i - X_j\|^\gamma\Big] \\ = \frac{2}{n}\mathbb{E}\|X_1 - X_2\|^\gamma.\end{aligned}$$

The conclusion then follows from taking the expectation of the expression

$$\mathbb{E}\Big[\mathcal{E}_\gamma^2(\tilde{\nu}_n, \nu_n)\Big|\nu_n\Big] = \mathcal{E}_\gamma^2(\nu,\nu_n) + \frac{1}{n}\mathbb{E}\|X_1 - X_2\|^\gamma$$

and the inequality $|x+y|^\gamma \le 2^{\max\{0,\gamma-1\}}(|x|^\gamma + |y|^\gamma)$ for all $x, y \in \mathbb{R}$. □

In other words, the expected energy distance between empirical and population measures decays at the parametric rate. This is not terribly surprising, as similar results hold for any MMD with bounded kernel. In addition, [85, Theorem 7] also shows using McDiarmid's inequality that the same quantity is sub-Gaussian with variance proxy $O(1/n)$.

Representation as a Sliced Distance

Another equivalent characterization of the generalized energy distance is as a "sliced" distance. Sliced distances are calculated by first choosing a random direction, and then computing a one-dimensional distance in the chosen direction between the projections of the two input distributions. One popular choice for the one-dimensional metric are the Wasserstein distances [133, 57, 132, 56, 158], as slicing alleviates the prohibitive computational burden due to the curse of dimensionality. Sliced distances are also studied for general one dimensional metrics [152, 131], and most relevantly for us, for Cramér's distance in particular [130, 129, 208]. The latter, as we shall see in Proposition 1.2.3 below, is precisely equivalent to the energy distance. To extend this beyond just the setting $\gamma = 1$, define the function

$$\psi_\gamma(x) = \begin{cases} |x|^{(\gamma-1)/2} & \text{for } \gamma \neq 1 \\ \mathbb{1}\{x \geq 0\} & \text{otherwise.} \end{cases} \tag{1.2.8}$$

To the best of our knowledge, the following result is novel for $\gamma \neq 1$. For the case $\gamma = 1$ see for example [190, Equation (2.5)].

Proposition 1.2.3. *Let $\gamma \in (0, 2)$ and let μ, ν be probability distributions on $\mathbb{R}^d$ with finite γ'th moment. Then for $(X, Y) \sim \mu \otimes \nu$ we have*

$$\mathcal{E}_\gamma^2(\mu, \nu) = \frac{1}{S_\gamma} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \Big[\mathbb{E}\psi_\gamma(\langle X, v\rangle - b) - \mathbb{E}\psi_\gamma(\langle Y, v\rangle - b) \Big]^2 \mathrm{d}b\mathrm{d}\sigma(v), \tag{1.2.9}$$

where $S_\gamma = \frac{\pi^{\frac{d}{2}+1}\Gamma(1-\frac{\gamma}{2})}{\gamma 2^{\gamma-1}\Gamma(\frac{d+\gamma}{2})\cos^2(\frac{\pi(\gamma-1)}{4})\Gamma(\frac{1-\gamma}{2})^2}$ *when* $\gamma \neq 1$ *and* $S_1 = \frac{\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d+1}{2})}$.

The proof of Proposition 1.2.3 hinges on computing the Fourier transform of the function ψ_γ, which can be interpreted as a tempered distribution. We point out a special property of the integral on the right hand side of (1.2.9): after expanding the square, one finds that the individual terms in the sum are not absolutely integrable for $\gamma \neq 1$. Nevertheless, due to cancellations within the squared quantity, the integral is finite.

1.3 Related work

In this section we survey prior results on problems related to the results of this thesis.

1.3.1 Estimation, Goodness-of-Fit and Two-Sample Testing

Ingster was the first to study minimax goodness-of-fit testing. In [112] he studies the problem for the Gaussian sequence model, see also [68]. In [111] he derives the minimax sample complexity of goodness-of-fit testing under L^2 separation for densities on $[0, 1]$ with bounded W_2^2 norm, where the W_β^q-norm denotes the L_q norm of the β'th derivative whenever β is an integer, and the usual generalization thereof for non-integers. In [110] he extends these results to L^p separated densities with bounded W_q^β norm for general $1 \leq p \leq \infty$ assuming $p \leq q$ when $p \leq 2$ and $p = q$ otherwise. In particular, setting $p = 1$ and $q = \infty$ recovers the

result from row $\mathcal{P}_{\mathsf{H}}$ and column n_{GoF} in Table 1.1. These results are extended to multiple dimensions and two-sample testing in [10, 139]. We are not aware of explicit prior work on two-sample testing for the Gaussian sequence model, but for regular cases such as the one we study, the results and methods carry over trivially. The estimation problems over $\mathcal{P}_{\mathsf{H}}$ and $\mathcal{P}_{\mathsf{G}}$ are even older, for the former see the work by Chentsov [43] Bretagnolle and Huber [33] and Ibragimov and Has'minskii [107] along with their book [109, Section IV.4]. For the latter class $\mathcal{P}_{\mathsf{G}}$, see Ibragimov and Has'minskii's book [109, Section VII.4] and their papers [108, 106].

	n_{HT}	n_{GoF}	$\mathcal{R}_{\mathsf{TS}}$	n_{Est}
$\mathcal{P}_{\mathsf{H}}(\beta, d)$	$1/\epsilon^2$	$1/\epsilon^{(2\beta+d/2)/\beta}$	$\min\{n, m\} \geq n_{\mathsf{GoF}}$	$\epsilon^2 n_{\mathsf{GoF}}^2$
$\mathcal{P}_{\mathsf{G}}(\beta, d)$	$1/\epsilon^2$	$1/\epsilon^{(2s+1/2)/s}$	$\min\{n, m\} \geq n_{\mathsf{GoF}}$	$\epsilon^2 n_{\mathsf{GoF}}^2$
$\mathcal{P}_{\mathsf{Db}}(\beta, d)$	$1/\epsilon^2$	$\sqrt{k}/\epsilon^2$	$\min\{n, m\} \geq n_{\mathsf{GoF}}$	$\epsilon^2 n_{\mathsf{GoF}}^2$
$\mathcal{P}_{\mathsf{D}}(\beta, d)$	$1/\epsilon^2$	$\sqrt{k}/\epsilon^2$	$\max\{n, m\} \geq \frac{\sqrt{k}}{\epsilon^2} + \frac{k^{2/3}}{\epsilon^{4/3}} \asymp n_{\mathsf{TS}}$ $\min\{n, m\} \geq n_{\mathsf{GoF}}\sqrt{\alpha}$	$\epsilon^2 n_{\mathsf{GoF}}^2$

Table 1.1: Prior results on testing and estimation.

Turning to the discrete case, we note that the survey [36] gives an excellent and highly detailed historical record of results and techniques. Goldreich and Ron introduced uniformity testing to the computer science literature in [81], and Batu et al. consider goodness-of-fit testing in [19]. In a sequence of follow up works, with [162] notably introducing the classical lower bound construction, the results were eventually improved to the optimal $\Theta(\sqrt{k}/\epsilon^2)$. We note that many of these results are in fact implied by Ingster [110]. Two-sample testing was solved in the milestone paper [199]. Estimating discrete distributions in total variation is solved optimally by the empirical measure. The first appearance of the result is unclear, the note [35] provides an exposition of related results. The aforementioned results all concerned the class $\mathcal{P}_{\mathsf{D}}$ of all discrete distributions, but one can easily specialize their proofs to apply to $\mathcal{P}_{\mathsf{Db}}$ and obtain optimal testers.

1.3.2 Likelihood-Free Hypothesis Testing

In this section we survey prior work relating to LFHT.

Ziv and Gutman

LHFT initially appeared in Gutman's paper [90], building on Ziv's work [212], where the problem is studied for distributions on a fixed, finite alphabet. Ziv called the problem "classification with empirically observed statistics", to emphasize the fact that hypotheses are specified only in terms of samples and the underlying true distributions are unknown. In

[209] it is shown that the error exponent of Gutman's test is second order optimal. Recent work [98, 93, 92, 30] extends this problem to distributed and sequential testing. However, the setting of these papers is fundamentally different from the one studied in this thesis, a point which we expand on below.

Given two arbitrary, unknown $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}$ over a finite alphabet of fixed size, Gutman's test (see [209, Equation (4)]) rejects the null hypothesis $H_0 : \mathbb{P}_{\mathsf{Z}} = \mathbb{P}_{\mathsf{X}}$ in favor of the alternative $H_1 : \mathbb{P}_{\mathsf{Z}} = \mathbb{P}_{\mathsf{Y}}$ if the statistic $\mathrm{GJS}(\widehat{\mathbb{P}}_{\mathsf{X}}, \widehat{\mathbb{P}}_{\mathsf{Z}}, \alpha)$ is large, where $\widehat{\mathbb{P}}$ denotes empirical measures, GJS denotes the generalized Jensen-Shannon divergence defined in [209, Equation (3)] and $\alpha = n/m$. In other words, it simply performs a two-sample test using the samples from $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Z}}$ of size n and m respectively, and completely discards the sample from $\mathbb{P}_{\mathsf{Y}}$. In light of our sample complexity results this is strictly sub-optimal due to minimax lower bounds on two-sample testing, see the difference of light gray and striped regions in Figure 1.1.

More generally, the method of types, which is a crucial tool for the works cited above, cannot be used to derive our results, because in the regime where the alphabet size k scales with the sample size n, the usual $\binom{n}{k} = e^{o(n)}$ approximation no longer holds, i.e. these factors affect estimation rates and do not lead to tight minimax results. As a consequence, one cannot deduce results about the minimax sample complexity of LFHT from works on the classical regime because the latter do not quantify the speed of convergence of the error terms as a function of the alphabet size. Specifically, let us examine [209, Theorem 1], which is a strengthening of the results of [90]. Paraphrasing, it states that for any fixed ratio $\alpha = n/m$ and pair of distributions $(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}})$, Gutman's test has type-II error bounded by $1/3$ when given samples from $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ as input, and type-I error bounded by $\exp(-\lambda n)$ given arbitrary input, where

$$\lambda = \mathrm{GJS}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \alpha) + \sqrt{\frac{V(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \alpha)}{n}}\Phi^{-1}(1/3) + \mathcal{O}\left(\frac{\log(n)}{n}\right) \tag{1}$$

as $n \to \infty$. Here V denotes the dispersion function defined in [209, Equation (9)] and Φ is the standard normal cdf. The crucial point we make here is that in (1) the dependence of the $\mathcal{O}(\log(n)/n)$ term on $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}$, and in particular their support size k and the ratio $\alpha = n/m$ is unspecified. Because of this, (1) and similar results cannot be used to derive minimax sample complexities as $\min\{n, m, k\} \to \infty$ jointly at possibly different rates.

This distinction between the fixed alphabet size setting studied in [90, 212, 209] and similar works, and our large alphabet setting was recognized by [102, 103, 123, 124] whose results are much closer to those presented in this thesis. In [103] Huang and Meyn introduce the concept of "generalized error exponent" to deal with support sizes that grow superlinearly with sample size (referred to as the "sparse sample regime" by them) in the setting of uniformity testing.[2] In [102] they extend this idea to LFHT and say, quote,

> "In the classification problem, the classical error exponent analysis has been applied to the case of fixed alphabet in [212] and [90].... However, in the sparse sample problem, the classical error exponent concept is again not applicable, and thus a different scaling is needed."

[2] Uniformity testing is the problem of goodness-of-fit testing where the null is given by a uniform distribution.

Kelly et al.

The first time that LFHT appeared formulated as a minimax problem is in [123, 124]; let us stick to the notation that we have introduced for LFHT. Both papers consider discrete distributions on the alphabet $\{1, 2, \dots, k\}$, and are the first to study LFHT in the minimax sense as n, m and k grow towards infinity. In more detail, these papers consider the setting when both P_X and P_Y (and consequently P_Z) are guaranteed to be close to uniform in the sense that the ratio of their pmfs to the uniform is bounded both above and below by a constant. This is simply our class $\mathcal{P}_{\mathsf{Db}}$ with an additional lower bound assumption. They assume moreover that the guaranteed separation, ϵ, between P_X and P_Y is fixed at a constant level and that $n = m$ (i.e. the "simulation" sample size and the real data sample size is the same). Finally, they are only concerned with characterizing when the sum of type-I and type-II error probabilities decays to 0, thus disregarding the dependence on the probability of error (δ in our notation) of the sample complexities.

The main result of the two papers is that the relation $n \gtrsim \sqrt{k}$ is necessary and sufficient in order to have vanishing probability of error [123, Theorem 3 & 4]. Due to the equivalence between LFHT and two-sample testing in the equal sample size $n = m$ case (see Proposition 1.4.1), their result follows from the fact that the minimax sample complexity of two-sample testing over their distribution class (which is a subset of our $\mathcal{P}_{\mathsf{Db}}$) is given by $\sqrt{k}/\epsilon^2$; setting $\epsilon = \Theta(1)$ recovers their claim. The test that they use for this is precisely the same as that in my first paper on LFHT [77], which we discuss in Section 1.4.

The results discussed so far are interesting on their own, however Kelly and his coauthors go on to show that some natural tests are not able to achieve this optimal sample complexity. Writing $\widehat{p}_X, \widehat{p}_Y, \widehat{p}_Z$ for the empirical pmfs of P_X, P_Y and P_Z respectively, each of the tests they consider are of the form:

$$\text{reject } H_0 \text{ if } d(\widehat{p}_X \| \widehat{p}_Z) \geq d(\widehat{p}_Y \| \widehat{p}_Z)$$

for some measure of distance d. Above we already mentioned that taking $d(a\|b) = \|a - b\|_2$ achieves the entire range $n \gtrsim \sqrt{k}$. Additional choices they consider are the Jensen-Shannon divergence, the χ^2-divergence and Hellinger distance.

They show that using either the Jensen-Shannon divergence [124, Theorem 3] or the χ^2-divergence [124, Theorem 5] in place of d results in vanishing error probability provided $n \gg k$ for arbitrary discrete distributions as input (i.e. the class $\mathcal{P}_{\mathsf{D}}$), but the performance breaks down as soon as $k \sim n$ even on the more regular class $\mathcal{P}_{\mathsf{Db}}$ [124, Theorem 4 & 6]. They conjecture, based on numerics, that the same negative result holds for the Hellinger distance.

Huang and Meyn

Follow-up work [102] extends the results of Kelly et al. to the case $m \neq n$, showing that LFHT is possible over the class $\mathcal{P}_{\mathsf{Db}}$ if and only if $k \lesssim \min\{n^2, nm\}$ [102, Theorem III.1.]. In addition, they go further and study the dependence on the probability of error δ in the 'sparse regime' $n, m = \mathcal{O}(k)$. They exhibit a novel test statistic based on collisions for which the sum of type-I and type-II decays as $\exp(-J \min\{n^2, nm\}/k)$ for some $J > 0$ (in the regime $\epsilon = \Theta(1)$ that is). Surprisingly, they also show that the difference of L^2-distances test analysed by Kelly et al. (and us later in this thesis) does not enjoy this rate i.e. $J = o(1)$.

Remark 4. *While the authors of [102] disregard the dependence on ϵ, their proof contains all that is necessary to derive the optimal dependence on ϵ when $m, n \leq k$; they even state after Proposition VI.1. that $J = \Theta(\epsilon^4)$. The optimal error decay over all of $\mathcal{P}_{\mathsf{Db}}$ reads $\exp(-\epsilon^4\Theta(\min\{nm/k, n^2/k, m/\epsilon^2\}))$, as we show in Section 1.5.*

Remark 5. *Huang's work [102, 103, 101] seems to have been somewhat underappreciated at the time. First, in the proof of [102, Theorem III.1.] he proposes to split each support element into multiple elements uniformly, an idea that form the basis of "flattening"/"Goldreich's reduction" [64, 82, 80, 62] and one that we use in our analysis of the class $\mathcal{P}_{\mathsf{D}}$ in some regimes. Second, he is the earliest reference that studies the dependence on the probability of error δ for both uniformity testing and LFHT. For both problems he identifies the optimal, sub-Gaussian dependence on δ in the sparse regime $n, m \leq k$, disregarding the dependence on the separation ϵ. Recent progress has resolved the high-probability sample complexity of goodness-of-fit testing [63] and two-sample testing [62], to which Huang's work is a precursor. Finally, [88] follows Huang's work (which uses analytical depoissonization [115] and the saddle point method [55]) closely to pin down down the minimax optimal sample complexity of uniformity testing up to a multiplicative $(1 + o(1))$ factor.*

Acharya et al.

Another related line of work studies LFHT (or 'classification' as the authors call it) in a competitive setting. [3, 2] considers discrete distributions in the case $n = m$ and measures sample complexity relative with respect to a symmetric oracle tester which has full knowledge of the underlying distributions P_X, P_Y but doesn't know whether $P_Z = P_X$ or $P_Z = P_Y$.[3] In [4] this problem is studied instead in the $m = 1$ case, i.e. the setting that is more classically referred to as classification in statistics and machine learning.

Nonparametric Classification

The last connection that we mention is to classification as studied in the nonparametric statistics community. Representative works are for example [205, 146], but the literature is vast and we don't attempt to review it here. In general, these papers are interested in the case $m = 1$ i.e. when we observe a sample of size 1 from P_Z and wish to assign it to either P_X or P_Y. Note that our LFHT problem is ill-posed when $m \ll 1/\epsilon^2$, and when $m = 1$ in particular, because even the optimal Neyman-Pearson likelihood-ratio test between P_X and P_Y, when fed a single observation from P_Z, will be wrong with probability $\frac{1}{2}(1 - \mathsf{TV}(P_X, P_Y))$ of the time. Thi optimal error probability is also called the "Bayes error". In contrast, in LFHT we require that this error be less than a constant, taken to be 0.3 throughout this thesis for concreteness. Therefore, these works have a different objective compared to ours: they study the rate at which the classification error approaches the Bayes error as a function of n. In [205] for example it is shown that this rate is equal to the rate of density estimation over many popular nonparametric classes, for example Besov spaces, with the takeaway that

[3] Symmetric means that the output of the test is measurable with respect to the joint fingerprint of the three samples from P_X, P_Y, P_Z. The joint fingerprint is simply the array with index $\mathbb{N}^3$ that for each triple (a, b, c) counts the number of support elements that receive a, b, c observations from P_X, P_Y, P_Z respectively. Note that the likelihood-ratio test is not symmetric in general.

"classification is no easier than estimating the conditional probability in a uniform sense". In the setting that we study this conclusion fails to hold, as LFHT turns out to be significantly "easier" than estimating the distributions P_X, P_Y and P_Z.

1.3.3 Classifier-Accuracy Testing

The first appearance of the idea of using a trained classifier to test whether two distributions are equal was in [72]. Since then, the method has enjoyed popularity in empirical circles, notably in neuroimaging [83, 210, 165] and likelihood-free inference [91, 67, 196, 5]. The theoretical understanding of the usage of classifiers for testing is much more limited. The only substantial work on the matter we are aware of is [126], where authors study the minimax power of classical algorithms (such as LDA) for Gaussian mean testing in high dimension. They also derive some results for general classifiers [126, Section 9], which apply exchangeability/permutation arguments or the central limit theorem to the classifier accuracy statistic (see (1.5.15) for the definition) as the test sample size goes to infinity, regarding the classifier as fixed.

1.4 LFHT in the Constant Error Regime

This section presents our first results about the region $\mathcal{R}_{\mathsf{LF}}$ defined in (1.2.2). More concretely, we fully characterize $\mathcal{R}_{\mathsf{LF}}$ in the constant error ($\delta = \Theta(1)$) regime for all classes $\mathcal{P}$ introduced in Section 1.2.2, with the exception of $\mathcal{P}_{\mathsf{D}}$, where we loose a $\log(k)$-factor. It is based on the preprint [77] which is joint work with Yury Polyanskiy, and is reproduced in full in Chapter 2. We start this chapter with some motivation for LFHT, and an outlook towards potential future research directions.

1.4.1 Motivation and Outlook

A setting called likelihood-free inference (LFI), also known as simulation based inference (SBI), has independently emerged in many areas of science over the past decades. Given an expensive to collect dataset and the ability to simulate from a high fidelity, often mechanistic, stochastic model, whose output distribution and likelihood is intractable and inapproximable, how does one perform model selection, parameter estimation or construct confidence sets? The list of disciplines where such highly complex black-box simulators are used is long, and include particle physics [9], astrophysics [7], climate science [97], epidemiology [155], neuroscience [70] and ecology [24] to just name a few. For some of the above fields, such as climate modeling, the bottleneck resource is in fact the simulated data as opposed to the experimental data. In either case, understanding the trade-off between the number of simulations and experiments required to do valid inference is crucial.

Let us make the above more formal. Suppose that we have a parameter set $\Theta \subseteq \mathbb{R}^p$ and a simulator $\mathcal{S} : \Theta \times [0,1] \to \mathbb{R}^d$, which given an external seed $U \sim \mathrm{Unif}([0,1])$ and a parameter setting $\theta \in \Theta$, produces the random variable $\mathcal{S}(\theta, U) \sim P_\theta$. In addition to the simulator $\mathcal{S}$, we also observe an i.i.d. dataset $Z_1, \ldots, Z_m$ with unknown distribution P_Z, that is produced by our real-world experiment. The goal of likelihood-free inference is to find a parameter

setting $\widehat{\theta}$ such that $P_{\widehat{\theta}}$ is close to P_Z according to some notion of similarity. Most often we will be using total variation, denoted TV, to measure distance between probability distributions. Instead of producing a good approximation $P_{\widehat{\theta}}$, one could also be interested in estimating the true parameter $\theta^* \in \arg\min_{\theta\in\Theta} \mathsf{TV}(P_\theta, P_Z)$ itself, or to construct confidence sets that are guaranteed to cover θ^* with certain probability.

In this thesis we are focused on minimax performance guarantees, that is, we search for algorithms that guarantee a certain level of performance given every input that satisfies a pre-specified constraint. It is intuitively obvious that the quality of performance we can guarantee will depend monotonically on the strength of these assumptions. Let us write $\mathcal{P}$ for a set of distributions that contains all potential simulator outputs, so that $\{P_\theta\}_{\theta\in\Theta} \subseteq \mathcal{P}$, and suppose that we have a finite number of possible parameter settings, say $\Theta = \{1, 2, \ldots, M\}$ for some $M \geq 2$. We outline some natural questions that arise from this model below.

Let us be given i.i.d. simulation samples from each of $P_1, \ldots, P_M$ of size $n_1, \ldots, n_M$ respectively. Write $i^* = \arg\min_{1\leq i\leq M} \mathsf{TV}(P_i, P_Z)$ for the parameter setting for which the simulator's output is closest to the real data distribution P_Z.

Question: How large must $\{n_i\}_{1\leq i\leq M}$ and m be to identify i^* w.h.p.? (Q1)

Here "w.h.p." abbreviates "with high probability". Clearly the answer to (Q1) will depend on the geometry of the set $\{P_i\}_{1\leq i\leq M}$. This chapter is about (Q1) in the case when $M = 2$, $P_Z \in \{P_1, P_2\}$ and under the assmption that $\mathsf{TV}(P_1, P_2) \geq \epsilon$ for some small $\epsilon > 0$. Recall from Section 1.2 that we call this problem "LFHT" which is short for "likelihood-free hypothesis testing". It is essentially a version of binary hypothesis testing where the two hypotheses P_1 and P_2 are specified only approximately via the two i.i.d. samples. An alternative and perhaps more apt name could have been "three-sample testing". As we'll see in our review of past work, the problem has also been called "classification" by the information theory literature. This is somewhat different from the modern usage of the word by the statistics and ML communities, where classification usually refers to the above problem in the special case $m = 1$.

Although what we have covered above is all we need to understand the results of this chapter, let us sketch some additional questions that could be the subject of future research. Readers can safely skip to Section 1.4.2.

Instead of requiring that we perfectly identify i^*, one could relax this to a list-decoding question. Let $\widehat{I}$ be a subset of $[M]$ that we choose based on the observed data.

Question: For given $\{n_i\}_{1\leq i\leq M}$ and m, how large must the set $\widehat{I}$ be to ensure that $i^* \in \widehat{I}$ w.h.p.? (Q2a)

Question (Q2a) is essentially the question of constructing confidence sets. Alternatively, if we are more interested in finding a parameter for which the simulator approximates the true distribution well, and don't care whether we approximate the best parameter itself closely, we can change (Q2a) to the following.

Question: How large must $\{n_i\}_{1\leq i\leq M}$ and m be to identify $\widehat{i}$ with $\mathsf{TV}(P_{\widehat{i}}, P_{i^*}) \leq \epsilon$ in expectation or w.h.p.? (Q2b)

A potential name for the problem (Q2b) could be "likelihood-free estimation" and can be interpreted as modeling a grid-search in LFI. In this interpretation we would run the simulator $n_1, \dots, n_M$ times with parameters $\theta_1, \dots, \theta_M$ respectively, producing i.i.d. datasets from the distributions $P_1 := P_{\theta_1}, \dots, P_M := P_{\theta_M}$ making up our grid-search.

If our resources are constrained, we might wish to minimize the total number of simulation samples $n_1, \dots, n_M$ used to identify our candidate $\widehat{i}$. Suppose that instead of drawing a pre-specified number n_i observations from P_i, we could choose an index $i \in [M]$ at each step and observe an independent draw from P_i; call such a sequential algorithm "active".

$$\textbf{Question: } \text{How large must } \textstyle\sum_{i=1}^{M} n_i \text{ and } m \text{ be to identify } \widehat{i} \text{ with } \mathsf{TV}(P_{\widehat{i}}, P_{i^*}) \le \epsilon \text{ in expectation or w.h.p. using an active algorithm?} \tag{Q3}$$

Actively choosing which parameters to simulate from is an important technique that is widely used by practitioners. For example, in the review [51] authors say that active learning is "...a key idea to improve the sample efficiency..." and that "even simple implementations can lead to a substantial improvement in sample efficiency".

1.4.2 General Reductions

Before diving into results specific to distribution classes, let me describe a number of general reductions between LFHT and the problems defined in Section 1.2.2. Some of these reductions are not with high probability. While sample splitting and majority voting can amplify any nontrivial success probability to, say $1-\delta$, this inflates the sample sizes by a factor of $\log(1/\delta)$ which we'll see is sub-optimal except for simple binary hypothesis testing. Therefore, some of the reductions below are only useful in the regime $\delta = \Theta(1)$, that is, when we only require a constant guarantee on the total error probability of our testing procedure.

Proposition 1.4.1. *Let $\mathcal{P}$ be a generic family of distributions. There exists a universal constant $c > 0$ such that for $n, m \in \mathbb{N}$ the following implications hold.*

$$(n,m) \in \mathcal{R}_{\mathsf{LF}} \overset{*}{\implies} m \ge n_{\mathsf{HT}}, \tag{1.4.1}$$

$$(n,m) \in \mathcal{R}_{\mathsf{TS}} \overset{*}{\implies} n \wedge m \ge n_{\mathsf{GoF}}, \tag{1.4.2}$$

$$(n,m) \in \mathcal{R}_{\mathsf{LF}} \implies cn \ge n_{\mathsf{GoF}}, \tag{1.4.3}$$

$$(n,m) \in \mathcal{R}_{\mathsf{TS}} \overset{*}{\implies} (n,m) \in \mathcal{R}_{\mathsf{LF}}, \tag{1.4.4}$$

$$m \ge n \text{ and } (n,m) \in \mathcal{R}_{\mathsf{LF}} \implies (cn, cm) \in \mathcal{R}_{\mathsf{TS}}, \tag{1.4.5}$$

where we omit the argument $(\epsilon, 0.3, \mathcal{P})$ for simplicity and implications marked with $*$ also hold with with the argument $(\epsilon, \delta, \mathcal{P})$. In particular

$$\mathbb{N}^2_{n \le m} \cap \mathcal{R}_{\mathsf{LF}}(\epsilon, 0.3, \mathcal{P}) \asymp \mathbb{N}^2_{n \le m} \cap \mathcal{R}_{\mathsf{TS}}(\epsilon, 0.3, \mathcal{P}), \tag{1.4.6}$$

where $\mathbb{N}^2_{n \le m} = \{(n,m) \in \mathbb{N}^2 : n \le m\}$.

Proof. In what follows, let $\Psi_{\mathsf{LF}}, \Psi_{\mathsf{TS}}$ be minimax optimal tests for LFHT and two-sample testing respectively. Throughout the proof we omit the arguments $(\epsilon, \delta, \mathcal{P})$ and $(\epsilon, 0.3, \mathcal{P})$ for notational simplicty.

Let us start by reducing hypothesis testing to LFHT. Suppose $(n, m) \in \mathcal{R}_{\mathsf{LF}}$. Let $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$ be given with $\mathsf{TV}(\mathbb{P}_0, \mathbb{P}_1) \geq \epsilon$ and suppose Z is an i.i.d. sample with m observations. We wish to test the hypothesis $H_0 : Z_i \sim \mathbb{P}_0$ against $H_1 : Z_i \sim \mathbb{P}_1$. To this end generate n i.i.d. observations X, Y from $\mathbb{P}_0, \mathbb{P}_1$ respectively, and simply output $\Psi_{\mathsf{LF}}(X, Y, Z)$. This shows that if $(n, m) \in \mathcal{R}_{\mathsf{LF}}$ then $m \geq n_{\mathsf{HT}}$ and concludes the proof of (1.4.1).

Next, we reduce goodness-of-fit testing to two-sample testing. Suppose $(n, m) \in \mathcal{R}_{\mathsf{TS}}$. Then obviously $(n \wedge m, \infty) \in \mathcal{R}_{\mathsf{TS}}$. However, two-sample testing with sample sizes $(n \wedge m, \infty)$ is equivalent to goodness-of-fit testing with a sample size of $n \wedge m$. Therefore, $n \wedge m \geq n_{\mathsf{GoF}}$ must hold, concluding the proof of (1.4.2).

Next, we reduce goodness-of-fit testing to LFHT. Suppose $(n, m) \in \mathcal{R}_{\mathsf{LF}}$ with $m \leq n$. Let a distribution $\mathbb{P}_0 \in \mathcal{P}$ be given as well as an i.i.d. sample X of size cn with unknown distribution P_X, where $c \in \mathbb{N}$ is a large integer. We want to test $H_0 : P_X = \mathbb{P}_0$ against $H_1 : P_X \in \mathcal{P}, \mathsf{TV}(P_X, \mathbb{P}_0) \geq \epsilon$. Generate $c \times 2$ i.i.d. samples $Y^{(i)}, Z^{(i)}$ for $i = 1, \ldots, c$ of size n, m respectively, all from $\mathbb{P}_0$. Split the sample X into c batches $X^{(i)}, i = 1, \ldots, c$ of size n each and form the variables

$$A_i = \Psi_{\mathsf{LF}}(X^{(i)}, Y^{(i)}, Z^{(i)}) - \Psi_{\mathsf{LF}}(X^{(i)}, Y^{(i)}, X^{(i+1)}_{1:m})$$

for $i = 1, 3, \ldots, 2\lfloor c/2 \rfloor - 1$, where $X^{(i)}_{1:m}$ denotes the first m observations in the batch $X^{(i)}$. Note that the A_i are i.i.d. and bounded random variables. Under the null hypothesis we have $\mathbb{E}A_i = 0$, while under the alternative they have mean $\mathbb{E}A_i \geq 0.3$. Therefore, a constant number $c/2$ observations suffice to decide whether $P_X = \mathbb{P}_0$ or not. In particular, $cn \geq n_{\mathsf{GoF}}$ which concludes the proof of (1.4.3) for the case $m \leq n$. The case $n \leq m$ follows from (1.4.5) and (1.4.2).

Now we reduce LFHT to two-sample testing. Suppose $(n, m) \in \mathcal{R}_{\mathsf{TS}}$. Let three samples X, Y, Z be given, of sizes a, a, b from the unknown distributions P_X, P_Y, P_Z respectively, where $\{a, b\} = \{n, m\}$. We want to test the hypothesis $H_0 : P_X = P_Z$ against $H_1 : P_Y = P_Z$. Then, the test

$$\widetilde{\Psi}_{\mathsf{LF}}(X, Y, Z) \coloneqq \Psi_{\mathsf{TS}}(X, Z)$$

shows that $(n, m), (m, n) \in \mathcal{R}_{\mathsf{LF}}$ and concludes the proof of (1.4.4).

We now reduce two-sample testing to LFHT. Suppose $(n, m) \in \mathcal{R}_{\mathsf{LF}}$ where $m \geq n$. Let two samples X, Y be given, from the unknown distributions $P_X, P_Y \in \mathcal{P}$ and of sample size cn, cm respectively, where $c \in \mathbb{N}$ is a large integer. We wish to test the hypothesis $H_0 : P_X = P_Y$ against $H_1 : \mathsf{TV}(P_X, P_Y) \geq \epsilon$. Split the samples X, Y into $2 \times c$ batches $X^{(i)}, Y^{(i)}, i = 1, \ldots, c$ of sizes n, m respectively, and form the variables

$$A_i = \Psi_{\mathsf{LF}}(X^{(i)}, Y^{(i)}_{1:n}, Y^{(i+1)}) - \Psi_{\mathsf{LF}}(Y^{(i)}_{1:n}, X^{(i)}, Y^{(i+1)})$$

for $i = 1, 3, \ldots, 2\lfloor c/2 \rfloor - 1$, where $Y^{(i)}_{1:n}$ denotes the first n observations in the batch $Y^{(i)}$. The variables A_i are i.i.d. and bounded. Under the null hypothesis we have $\mathbb{E}A_i = 0$ while under the alternative $\mathbb{E}A_i \geq 0.3$ holds. Therefore a constant number $c/2$ observations suffice to decide whether $P_X = P_Y$ or not. In particular, $(cn, cm) \in \mathcal{R}_{\mathsf{TS}}$ which concludes the proof of (1.4.5).

Finally, for the equivalence between two-sample testing and LFHT, namely Equation (1.4.6), follows immediately from (1.4.5) and (1.4.4). □

All reductions above, with the possible exception of Equation (1.4.6), are obvious. Equation (1.4.6) says that the problems of likelihood-free hypothesis testing and two-sample testing are equivalent for $m \geq n$, i.e. when we have more real data than simulated data. We will see in the next section (and on Figures 1.1 and 1.3 visually) that the distinction between $n \leq m$ and $m \leq n$ is necessary.

1.4.3 Ingster's Goodness-of-Fit Test for Smooth Distributions

In his seminal paper [110] Ingster computes the minimax sample complexity of uniformity testing for a range of Sobolev spaces, measuring separation with respect to $L^p, p \geq 1$. Additionally, he provides matching minimax lower bounds for each of the cases he considers. Most relevant to us, he obtains the famous formula

$$n_{\mathsf{GoF}}(\epsilon, \delta = \Theta(1), \mathcal{P}_{\mathsf{H}}(\beta, d = 1, C = \Theta(1))) \asymp \epsilon^{-(2\beta+1/2)/\beta}.$$

This result stands in contrast with the, at that time already known, density estimation result $n_{\mathsf{Est}} \asymp \epsilon^{-(2\beta+1)/\beta}$: the sample complexity of goodness-of-fit testing enjoys a savings of $1/(2\beta)$ in the exponent of ϵ. This fact, that testing is (substantially) easier than estimation, is more general and continues to hold in general dimension d in which case sample complexities read $\epsilon^{-(2\beta+d/2)/\beta}$ [10, 139] and $\epsilon^{-(2\beta+d)/\beta}$ respectively.

Ingster's approach was to reduce to the problem of uniformity testing over the class $\mathcal{P}_{\mathsf{Db}}$ (bounded discrete distributions) by discretizing the interval $[0, 1]$ using a regular grid. Once discretized, he simply computes the L^2-distance between the known null distribution's p.m.f. and the empirical p.m.f. of the data. More concretely, given i.i.d. data $X_1, \ldots, X_n \in [0, 1]$ and a resolution k, form the empirical p.m.f.

$$\widehat{p}_i = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\left\{ X_j \in [\frac{i-1}{k}, \frac{i}{k}] \right\},$$

and reject the null hypothesis that the X_i are drawn from the uniform distribution on $[0, 1]$ if

$$\|\widehat{p}_i - \mathbb{1}/k\|_2 \geq \gamma$$

for an appropriate (k, n)-dependent threshold γ.

Notice that $\|\widehat{p}_i - \mathbb{1}/k\|_2^2 = \sum_{i=1}^k \widehat{p}_i^2 - 1/k$, so Ingster's test for uniformity is equivalent to thresholding $\sum_{i=1}^k \widehat{p}_i^2$, which is known as the collision statistic. The name comes from the fact that $\sum_{i=1}^k \widehat{p}_i^2 = \#\{(i, j) : X_i = X_j\}/n^2$. While largely overlooked by the computer science community, Ingster's work subsumes a lot of early progress in the distribution testing literature [81, 20], including the upper and lower bound analyses as well as the test statistic itself.

The key step enabling Ingster's reduction is to show that the separation between any two smooth distributions doesn't degrade too much by going from continuous to discrete.

Lemma 1.4.2 ([10, Lemma 7.2]). *Given $r \in \mathbb{N}$ and $j \in \{1, \ldots, r\}^d$, let $B_j = (j - \mathbb{1})/r + [0, 1]^d/r$ be the hypercube of sidelength $1/r$ with center $(j - \mathbb{1}/2)/r$. Define the map P_r by*

$$(P_r f)(x) = \sum_{j \in [r]^d} \mathbb{1}\{x \in B_j\} \left(\frac{1}{1/r^d} \int_{B_j} f(y) \mathrm{d}y \right).$$

Then, for any $\beta, C > 0$ and $d \geq 1$ there exist constants $c, c' > 0$ such that for any $f \in \mathcal{C}(\beta, d, C)$ we have

$$\|f\|_2 \geq \|P_r f\|_2 \geq c\|f\|_2 - c' r^{-\beta}.$$

Remark 6. *Note that P_r is the L^2 projection onto the space of functions piecewise constant on the cubes B_j, which assigns for each $x \in B_j$ the average value of the input on the box B_j. We can also regard P_r as a projector onto the Haar wavelet basis.*

If we take f to be a probability density in Lemma 1.4.2, then $P_r f$ is precisely the density of the discretization of the original distribution on the regular grid. Therefore, Lemma 1.4.2 can be used to justify the reduction of Ingster from $\mathcal{P}_{\mathsf{H}}$ to $\mathcal{P}_{\mathsf{Db}}$: taking $r \asymp \epsilon^{-1/\beta}$ preserves ϵ-separation with respect to L^2 distance, which can then be converted to separation under total variation via Hölder's inequality.

On first look it might be surprising that Lemma 1.4.2 is true and can lead to minimax optimal testing results for all levels of smoothness $\beta > 0$. Suppose we observe i.i.d. observations from the unknown β-smooth density f on $[0,1]^d$ and consider the following scheme to estimate f: simply output the histogram of the observations when discretized on the regular grid with mesh $r = \epsilon^{-1/\beta}$. It is well known (see for example [35, Theorem 1]) that we need $\mathcal{O}((r^d + \log(1/\delta))/\epsilon^2)$ observations to estimate any discrete distribution with support size r^d within ϵ-TV distance with probability $1 - \delta$. Therefore, if we take $r = \epsilon^{-1/\beta}$, this process seems to produce an estimator of f with the minimax optimal sample complexity $\epsilon^{-(2\beta+d)/\beta}$. However, it is common knowledge that this is not true for $\beta > 1$!

Indeed, to complete our fictitious analysis of the proposed density estimator we need an upper bound on $\|f - P_r f\|_2$ of order $r^{-\beta}$, which can be shown to be false in general. The best such result that holds in general is $\|f - P_r f\|_2 \lesssim r^{-\beta \wedge 1}$, and this can be seen from the following simple example. Suppose that $f(x) = 1/2 + x$ for $x \in [0,1]$ so that f lies in our β-smooth class $\mathcal{P}_{\mathsf{H}}$ for any $\beta > 0$. Suppose we discretize f on a grid with mesh $1/r$ using the operator P_r defined in Lemma 1.4.2. Then

$$\|f - P_r f\|_2^2 = r \int_{-1/(2r)}^{1/(2r)} x^2 \mathrm{d}x = \frac{1}{12r^2},$$

which proves that $\|f - P_r f\|_2 \lesssim r^{-1}$ is the best bound one can hope for even for $\beta \geq 1$.

Summarizing, taking histograms produces optimal density estimators only for $\beta \leq 1$. This does not contradict our results in the previous subsection, because Lemma 1.4.2 only asserts that $\|f\|_2 \approx \|P_r f\|_2$ whenever $\|f\|_2 \gtrsim r^{-\beta}$ and makes no claims about the magnitude of $\|f - P_r f\|_2$.

1.4.4 Results for Regular Classes

Inspired by Ingster, we proposed the following test for LFHT.[4] To simplify our presentation, let us focus on the case of $\mathcal{P}_{\mathsf{Db}}$, that is, bounded discrete distributions with support in $\{1, 2, \ldots, k\}$. We observe i.i.d. random variables $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$ and $Z_1, \ldots, Z_m$ from

[4]Note that this statistic was already studied by [123, 124, 102] in some regimes, see Section 1.3.2 for more on the history.

P_X, P_Y, P_Z respectively. Writing p_X, p_Y, p_Z for the corresponding p.m.f.s, we construct their empirical versions, that is, the normalized empirical frequencies denoted $\widehat{p}_X, \widehat{p}_Y, \widehat{p}_Z$. Our test

$$\text{rejects the null if } \|\widehat{p}_X - \widehat{p}_Z\|_2 \geq \|\widehat{p}_Y - \widehat{p}_Z\|_2 \text{ and accepts it otherwise.}$$

A conceptually simple, although long, mean-variance analysis of this statistic coupled with Chebyshev's inequality yields the following result characterizing $\mathcal{R}_{\mathsf{LF}}$ for the three "regular" classes and when $\delta = \Theta(1)$.

Theorem 1.4.3. *For each choice $\mathcal{P} \in \{\mathcal{P}_{\mathsf{H}}, \mathcal{P}_{\mathsf{G}}, \mathcal{P}_{\mathsf{Db}}\}$ we have*

$$\mathcal{R}_{\mathsf{LF}}(\epsilon, 0.3, \mathcal{P}) \asymp \left\{ m \geq 1/\epsilon^2, n \geq n_{\mathsf{GoF}}(\epsilon, 0.3, \mathcal{P}), mn \geq n_{\mathsf{GoF}}(\epsilon, 0.3, \mathcal{P})^2 \right\},$$

where the implied constants do not depend on k (in the case of $\mathcal{P}_{\mathsf{Db}}$) or ϵ.

While we won't cover the tedious calculation here, let us point out one aspect of the analysis. Notice that the region that we obtain $\mathcal{R}_{\mathsf{LF}}$ is asymmetric, since m can be as small as $\mathcal{O}(1/\epsilon^2)$, but n is constrained to be at least $n_{\mathsf{GoF}} \gg 1/\epsilon^2$. As mentioned above, the achievability direction of Theorem 1.4.3 follows by analyzing the mean and variance of the statistic

$$\begin{aligned} T &= \|\widehat{p}_X - \widehat{p}_Z\|_2^2 - \|\widehat{p}_Y - \widehat{p}_Z\|_2^2 \\ &= \frac{\|\widehat{p}_X\|_2^2 - \|\widehat{p}_Y\|_2^2}{n^2} + \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}(\widehat{p}_Y(i) - \widehat{p}_X(i))\widehat{p}_Z(j). \end{aligned}$$

From the expanded expression in the second line above one can already glean where the conditions on n and mn may come from. The key point is that the $\|\widehat{p}_Z\|_2^2/m^2$ term is cancelled, which is precisely what allows us to take $1/\epsilon^2 \lesssim m \ll n_{\mathsf{GoF}}$. Phrased another way, the values of $\|\widehat{p}_X - \widehat{p}_Z\|_2$ and $\|\widehat{p}_Y - \widehat{p}_Z\|_2$ are awful as estimators of the population distances $\|p_X - p_Z\|_2$ and $\|p_Y - p_Z\|_2$; however they are "awful in the same direction". We defer the discussion of the lower bound constructions to Section 1.7

Figure 1.1 visualizes the region $\mathcal{R}_{\mathsf{LF}}$ for the classes covered in Theorem 1.4.3, and Figure 1.2 shows the empirical trade-offs for LFHT and two-sample testing from simulation on a toy problem.[5] In light of the reductions in Proposition 1.4.1, we see that the points $\{\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}\}$ of Figure 1.1 have special interpretations. A corresponds to the limit as $n \to \infty$. In this case LFHT is equivalent to binary hypothesis testing in which case $m = \Theta(1/\epsilon^2)$ observations suffice from P_Z. On the other extreme, D shows the limit as $m \to \infty$. In this case we know the distributions P_Z exactly, but we do not know whether it is equal to P_X or P_Y. Discarding, say, the Y-sample and applying an optimal goodness-of-fit testing procedure to test whether $P_X = P_Z$, we can prove that the point $\mathsf{D} = (m = \infty, n = n_{\mathsf{GoF}})$ is achievable. Conversely, our reduction (1.4.3) shows that this is the best possible, that is, n cannot decrease below n_{GoF}.

The lines $m = 1/\epsilon^2$ and $nm = n_{\mathsf{GoF}}^2$ intersect at $\mathsf{B} = (n_{\mathsf{GoF}}^2\epsilon^2, 1/\epsilon^2)$, this is the point with the minimal possible value of n for which $m = \Theta(1/\epsilon^2)$ still suffices. As can be seen from

[5] In this toy problem we set $\epsilon = .3, k = 100$ and $P_X = P_Z, P_Y$ are distributions on $\{1, 2, \ldots, k\}$ with $P_X(i) = (1 + \epsilon \cdot (2 \cdot \mathbb{1}\{i \text{ odd}\} - 1))/k = 2/k - P_Y(i)$ for all $i = 1, 2, \ldots, k$.

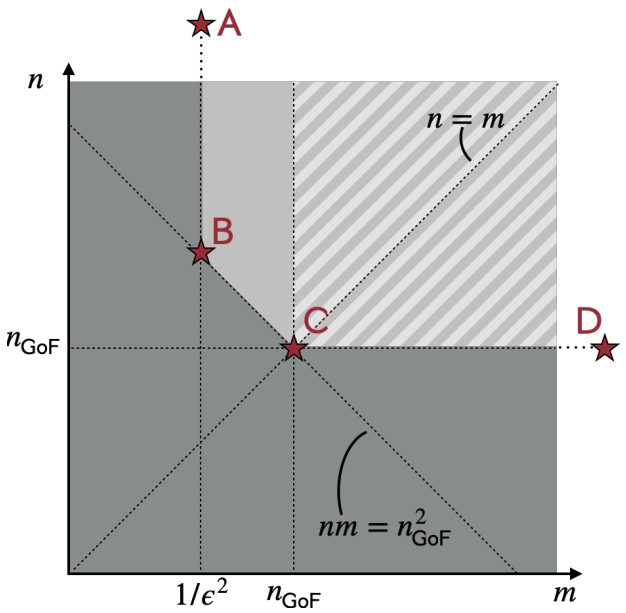
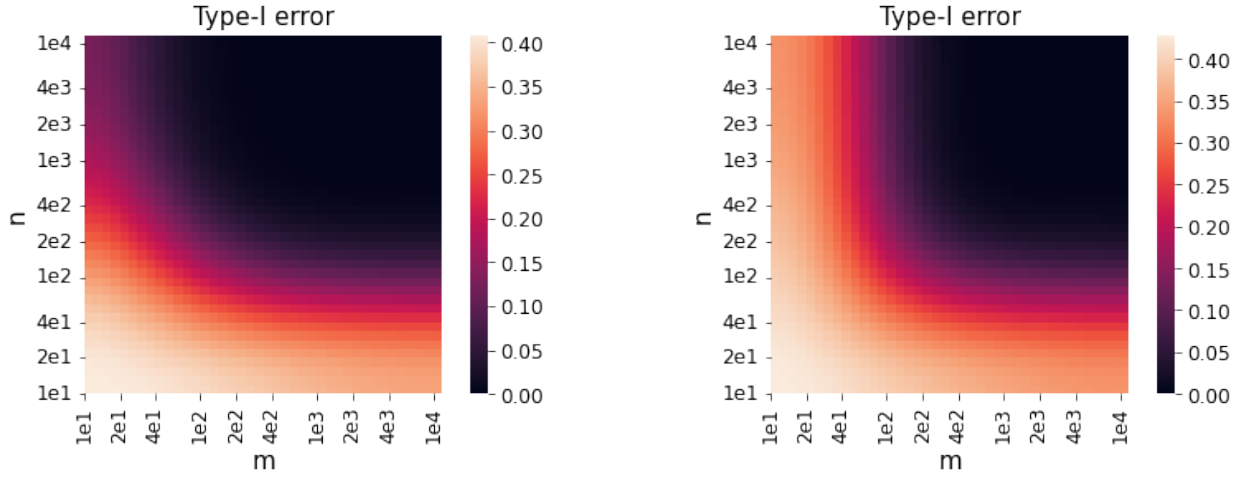


Figure 1.1: log-scale plot of region where LFHT is possible at constant error level $\delta = \Theta(1)$ as per Theorem 1.4.3.



(a) Likelihood-free hypothesis testing.

(b) Two-sample testing.

Figure 1.2: n versus m trade-off for a toy experiment. Probabilities estimated over 10^4 runs, and smoothed using Gaussian noise.

Table 1.1, it turns out that $n_{\mathsf{GoF}}^2\epsilon^2$ is equal up to constant to the rate of estimation n_{Est} for all four distribution classes we consider. With this perspective the achievability of the point $\mathsf{B} = (n_{\mathsf{Est}}, 1/\epsilon^2)$ is obvious: if we collect enough observations from P_X and P_Y to estimate them to high accuracy, that is, to within a small fraction of the separation between the two, then we can simply perform an ordinary robust binary hypothesis test between $\widehat{P}_X, \widehat{P}_Y$ using the m observations from P_Z.[6]

[6]For a TV-robust hypothesis test see the seminal [104] and for Hellinger see [27].

Finally, note from Table 1.1 that $n_{\mathsf{GoF}} = n_{\mathsf{TS}}$ for the classes considered in Theorem 1.4.3. We saw in Proposition 1.4.1 that two-sample testing and LFHT with $n = m$ are equivalent, and it is this point on the trade-off that $\mathsf{C} = (n_{\mathsf{GoF}}, n_{\mathsf{GoF}}) = (n_{\mathsf{TS}}, n_{\mathsf{TS}})$ represents.

The sample complexity of LFHT naturally interpolates between that of multiple foundational statistical problems. We point out a curious fact that is obvious as a result of the above discussion: since the product of n and m remains constant on the line segment $[\mathsf{B}, \mathsf{C}]$ on the left plot of Figure 1.1, it follows that

$$n_{\mathsf{Est}}(\epsilon, 0.3, \mathcal{P}) \asymp n^2_{\mathsf{GoF}}(\epsilon, 0.3, \mathcal{P})\, \epsilon^2 \tag{1.4.7}$$

for each class $\mathcal{P}$ treated in Theorem 1.4.3. This relation between the sample complexity of estimation and goodness-of-fit testing has not been observed before to our knowledge, and the generality of this phenomenon remains open.

Question: For what classes $\mathcal{P}$ does the relation (1.4.7) hold? (Q5)

It seems likely that the answer is along the following lines: classes for which a mixture over a hypercube provides the least favorable construction. According to personal correspondence with Zeyu Jia and Yury Polyanskiy, there exist convex bodies for which (1.4.7) breaks down under the Gaussian sequence model.

1.4.5 Results for the Unrestricted Discrete Class

Theorem 1.4.3 doesn't cover the class $\mathcal{P}_{\mathsf{D}}$ of all discrete distributions. This is no accident, as this case requires special attention. The proof of Theorem 1.4.3, that is, the analysis of the L^2-based test utilizes the fact that the density is bounded repeatedly. However, such a bound is no longer available in the case of $\mathcal{P}_{\mathsf{D}}$.

In order to salvage the analysis from the regular cases, we searched for a way to reduce from $\mathcal{P}_{\mathsf{D}}$ back to $\mathcal{P}_{\mathsf{Db}}$ or at least to obtain some control over the infinity norm of the probability mass functions. It turns out that in the distribution testing literature this reduction is known as "flattening" [102, 64, 80]. The idea is to split your dataset into two parts. Using the empirical frequencies from the first part you obtain an estimate of the mass placed on each element of the support. Then, you divide each support element into additional artificial buckets whose number is proportional to the empirical frequency in the first split. This processing step maintains total variation distances and thus the ϵ-separation between P_X and P_Y. Furthermore, using a union bound where we lose a factor of $\log(k)$, this ensures that the maximum of the p.m.f.s of both distributions P_X, P_Y are controlled. Then, on the second split of the data we simply apply the L^2-based test just as before. This strategy leads us to the following result.

Theorem 1.4.4. *Let $\alpha = 1 \vee (\frac{k}{n} \wedge \frac{k}{m})$. Then*

$$\mathcal{R}_{\mathsf{LF}}(\epsilon, 0.3, \mathcal{P}_{\mathsf{D}}) \asymp_{\log(k)} \begin{Bmatrix} m \geq 1/\epsilon^2, n \geq n_{\mathsf{GoF}}(\epsilon, 0.3, \mathcal{P}_{\mathsf{D}}) \cdot \sqrt{\alpha} \\ mn \geq n_{\mathsf{GoF}}(\epsilon, 0.3, \mathcal{P}_{\mathsf{D}})^2 \cdot \alpha \end{Bmatrix},$$

where the equivalence is up to a logarithmic factor in the alphabet size k.

Remark 7. *Note that* $\mathcal{R}_{\mathsf{LF}}(\epsilon, 0.3, \mathcal{P}_{\mathsf{D}}(k)) \asymp \mathcal{R}_{\mathsf{LF}}(\epsilon, 0.3, \mathcal{P}_{\mathsf{Db}}(k))$ *in the regime* $k < 1/\epsilon^4$.

See Figure 1.3 for a depiction of the region in Theorem 1.4.4. Our follow-up work, which we cover in Section 1.5, along with prior work on two-sample testing [62] removes the $\log(k)$-factor in Theorem 1.4.4.

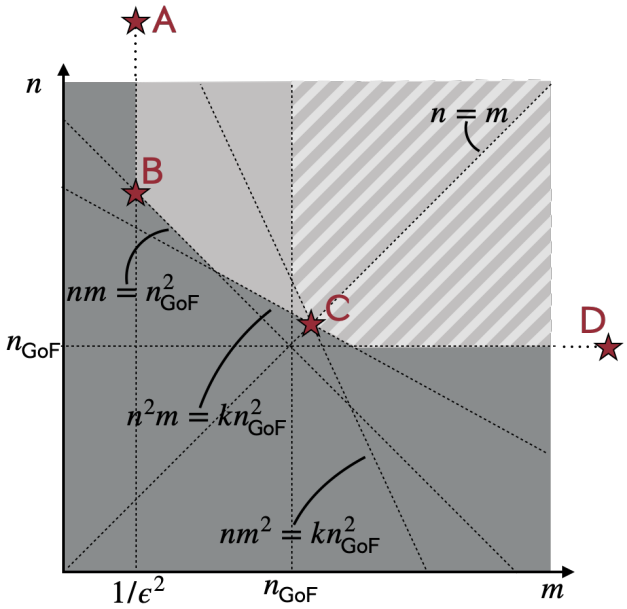


Figure 1.3: log-scale plot of region where LFHT is possible at constant error level over $\mathcal{P}_{\mathsf{D}}$ in the regime $k \geq 1/\epsilon^4$. For $k < 1/\epsilon^4$ refer to Figure 1.1.

1.5 Testing and Estimation by Classification

In this chapter we study testing and estimation algorithms whose key step is identifying sets, or equivalently, binary classifiers, where two distributions differ.

In Sections 1.5.1 to 1.5.3 we look at using half-spaces to separate distributions, and how our insights can be used for density estimation and two-sample testing. Most of this material is based on the preprint [75] which is joint work with Tianze Jiang, Yury Polyanskiy and Rui Sun, which can be found in full detail in Chapter 5.

Second, in Sections 1.5.4 and 1.5.5 we will look at learning arbitrary separating sets, which will lead, among other results, to achieving the optimal high probability sample complexity of LFHT, improving on our results from Section 1.4. This work is based on [74] which was published at the 36th Annual Conference on Learning Theory and is joint work with Yanjun Han and Yury Polyanskiy, and can be found in full detail in Chapter 4.

1.5.1 Separating Distributions by Half-Spaces

We define the "half-space discrepancy" $\overline{d_H}$ between two distributions μ and ν as

$$\overline{d_H}(\mu, \nu) = \sup_{\text{half-spaces } \Sigma} \big\{\nu(\Sigma) - \nu(\Sigma)\big\}.$$

In other words, it is the maximal difference in mass that μ and ν place on any half-space. We start with a simple, but important observation.

Lemma 1.5.1. *$\overline{d_H}$ defines a metric over the space of probability measures.*

Proof. Clearly, for any probability measures μ, ν and σ, it holds that $\overline{d_H}(\mu,\nu) = \overline{d_H}(\nu,\mu)$, $\overline{d_H}(\mu,\mu) = 0$ and $\overline{d_H}(\mu,\nu) \leq \overline{d_H}(\mu,\sigma) + \overline{d_H}(\sigma,\nu)$. The last property that we must prove to conclude is that $\overline{d_H}(\mu,\nu) = 0$ implies that $\mu = \nu$. For this last step, we follow the argument in the proof of [161, Theorem 4], which introdced $\overline{d_H}$ concurrently to our work. They study $\overline{d_H}$ and generalizations of it for two-sample testing, we comment on their results more extensively in Section 1.5.3. If $X \sim \mu$ and $Y \sim \nu$ with $\mu \neq \nu$, there must exist $\omega \in \mathbb{R}^d$ with

$$\langle \omega, X \rangle \overset{d}{\neq} \langle \omega, Y \rangle,$$

as otherwise the Fourier transforms $\widehat{\mu}$ and $\widehat{\nu}$ would coincide. The conclusion follows by the Fourier inversion theorem. □

Recalling that $\mathcal{E}_1$ denotes the energy distance, by Proposition 1.2.3 we have that

$$\mathcal{E}_1^2(\mu,\nu) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi^{\frac{d-1}{2}}} \underbrace{\int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \Big(\mathbb{P}(\langle X, v\rangle \geq b) - \mathbb{P}(\langle Y, v\rangle \geq b)\Big)^2 \mathrm{d}b \mathrm{d}\sigma(v)}_{=:d_H},$$

where $\mathrm{d}\sigma$ is the surface measure on $\mathbb{S}^{d-1}$, and $X \sim \mu$ and $Y \sim \nu$. Clearly d_H is similar to $\overline{d_H}$, except it takes an *average* over half-spaces instead of a maximum. This analogy isn't quite precise, as we integrate b over the entire real line, therefore it is not a bona fide average and the inequality $d_H \lesssim \overline{d_H}$ doesn't have to hold in general. In the next few sections we show that such an inequality does hold for certain structured classes of distributions, and show that distributions belonging to these classes can be separated by half-spaces surprisingly well.

Smooth Distributions

In Section 1.2.3 we already saw the class $\mathcal{P}_{\mathsf{H}}(\beta, d, C)$ of densities on $[0,1]^d$ whose β'th derivative is bounded. Here we consider a related class. Given $f \in L^2(\mathbb{R}^d)$ and $\beta > 0$, define the homogenous Sobolev (semi)norm of order $(\beta, 2)$ of f as

$$\|f\|_{\beta,2}^2 := \int_{\mathbb{R}^d} \|\omega\|^{2\beta} |\widehat{f}(\omega)|^2 \mathrm{d}\omega.$$

When β is an integer, $\|f\|_{\beta,2}$ is equal, up to constant, to the L^2-norm of the β'th derivative of f. We define $\mathcal{P}_S(\beta, d, C)$ to be the set of distributions on $\mathbb{R}^d$ that have density p with $\mathrm{supp}(p) \subseteq \mathbb{B}(0,1)$ and $\|p\|_{\beta,2} \leq C$, where $\mathbb{B}(x,r)$ is the ball centered at x with radius r. Going forward, to avoid trivialities, we assume that C is large enough in terms of β and d so that $\mathcal{P}_S(\beta, d, C)$ contains infinitely many distributions.

Lemma 1.5.2. *For every $\beta > 0, d \geq 1$ and $C > 0$ there exists a finite constant c such that for any $f, g \in \mathcal{P}_S(\beta, d, C)$ it holds that*

$$\mathsf{TV}(f,g)^{\frac{2\beta+d+1}{2\beta}} \leq c\, \overline{d_H}(f,g). \tag{1.5.1}$$

Proof. Let us write $\|g\|_{t,2}^2 := \int_{\mathbb{R}^d} |\widehat{g}(\omega)|^2 \|\omega\|^{2t} \mathrm{d}\omega$ for $g \in L^2$ and $t \in \mathbb{R}$. In the inequalities below we use the symbols $\lesssim, \asymp$ freely, indicating inequalities that hold up to constants involving d, β and C. Given any $\varphi > 0$ and Hölder conjugates r, r^*, we have

$$\|f-g\|_2^2 \overset{\text{Parseval}}{\asymp} \|\widehat{f}-\widehat{g}\|_2^2 \overset{\text{Hölder}}{\leq} \underbrace{\|f-g\|_{\frac{\varphi r}{2},2}^{2/r}}_{\leq C^{2/r}} \|f-g\|_{-\frac{\varphi r^*}{2},2}^{2/r^*}. \tag{1.5.2}$$

Choosing φ and r to satisfy $\varphi r = 2\beta$ and $\varphi r^* = d+1$, we get the chain of inequalities

$$\mathsf{TV}(f,g) \overset{\text{Jensen}}{\lesssim} \|f-g\|_2 \overset{\text{Eq. (1.5.2)}}{\lesssim} \left(\|f-g\|_{-\frac{d+1}{2},2}\right)^{\frac{2\beta+d+1}{2\beta}} \overset{\text{Prop. 1.2.1}}{\asymp} \left(\mathcal{E}_1(f,g)\right)^{\frac{2\beta}{2\beta+d+1}}.$$

The final step is to note that $d_H(f,g) \lesssim \overline{d_H}(f,g)$ due to the compact support of f and g. □

In the proof above we used two inequalities: Jensen's inequality to pass from TV to the L^2 distance, and Hölder's inequality in (1.5.2) to bound the L^2 distance between f and g by the energy distance times the homogenous Sobolev norm of their difference. Therefore, in order to prove that (1.5.1) is the best possible result one can obtain in general, we just have to exhibit a construction that saturates both inequalities at the same time.

In order for our application of Hölder's inequality to be tight, the equality

$$\frac{|\widehat{f}(\omega)-\widehat{g}(\omega)|^2}{\|\omega\|^{d+1}} = c|\widehat{f}(\omega)-\widehat{g}(\omega)|^2\|\omega\|^{2\beta}$$

would have to hold for almost every $\omega \in \mathbb{R}^d$ and some fixed c. This is clearly only possible if $\widehat{f}-\widehat{g}$ is supported on a sphere of some radius k centered at 0, in which case $c = (1/k)^{2\beta+d+1}$. Unfortunately, this is impossible for $f, g \in \mathcal{P}_S(\beta, d, C)$, because it would contradict the requirement of compact support. Nevertheless, we are able to use this idea, and take $f-g$ to be a suitably modified version of the inverse Fourier transform of a sphere (which is a Bessel function of the first kind [203]), leading to the following result.

Lemma 1.5.3 ([75, Proposition 7])**.** *For any $\beta > 0$, $d \geq 1$ and large enough $C > 0$, there exists a finite constant C_1 so that for any value of $\epsilon \in (0,1)$, there exist $\mu_\epsilon, \nu_\epsilon \in \mathcal{P}_S(\beta, d, C)$ such that $\mathsf{TV}(\mu_\epsilon, \nu_\epsilon)/\epsilon \in (1/C_1, C_1)$ and*

$$\overline{d_H}(\mu_\epsilon, \nu_\epsilon) \leq C_1 \mathsf{TV}(\mu_\epsilon, \nu_\epsilon)^{\frac{2\beta+d+1}{2\beta}} \log\left(\frac{3}{\mathsf{TV}(\mu_\epsilon, \nu_\epsilon)}\right)^{d-1}.$$

The bound in Lemma 1.5.3 above matches our bound from Lemma 1.5.2 exactly in one dimension, and up to a logarithmic factor in dimension two and above.

Gaussian Mixtures

Let us write

$$\mathcal{P}_G(d) := \{\nu * \mathcal{N}(0, I_d) : \nu \in \mathcal{P}(\mathbb{R}^d), \mathrm{supp}(\nu) \subseteq \mathbb{B}(0,1)\} \tag{1.5.3}$$

for the set of all Gaussian mixtures with support in the unit ball. We are able to derive a result similar to Lemma 1.5.2 for this class of distributions too.

Lemma 1.5.4. *For any $d \geq 1$ there exists a finite constant c such that for any $f, g \in \mathcal{P}_G(d)$ it holds that*

$$\frac{\mathsf{TV}(f,g)}{\log(3/\mathsf{TV}(f,g))^{\frac{d+1}{2}}} \leq c\,\overline{d_H}(f,g).$$

Proof sketch. The equation (1.5.2) still holds, except now we may let $\varphi r =: 2\beta$ be as large as we wish. Note that $\|f-g\|_{\frac{\varphi r}{2},2}$ can be bounded by the 2β'th moment of a Gaussian distribution. Optimizing over β we find that the best choice is given by $\beta \asymp \log(1/\|f-g\|_2)$.

There are two additional steps to the proof, which involve bounding $\mathsf{TV}(f,g)$ as a function of $\|f-g\|_2$ and d_H as a function of $\overline{d_H}$. The former follows by a straightforward generalization of [116, Theorem 22]. The latter follows by decomposing the integral in the definition of d_H into integration over a centered ball of radius $\sqrt{\log(1/d_H(f,g))}$ and its complement. □

In other words, given any two Gaussian mixtures with total variation separation ϵ, there always exists a half-space on which their mass differs by $\tilde{O}(\epsilon)$.

Discrete Distributions

Suppose we have two discrete distributions that are supported on a common, finite set of size k. One way to measure the energy distance between them would be to identify their support with the set $\{1, 2, \ldots, k\}$, thereby embedding the two distributions in $\mathbb{R}$, and applying the one-dimensional energy distance.

While the above approach seems reasonable, it is entirely arbitrary. Indeed, there might not be a natural ordering of the support; moreover, why should one choose the integers between 1 and k instead of, say, the set $\{1, 2, 4, \ldots, 2^k\}$? The total variation distance does not suffer from such ambiguities, and it is unclear how our choice of embedding affects the relationship to TV. The following result attacks precisely this question.

Theorem 1.5.5. *Let μ and ν be probability distributions supported on the set $\{x_1, \ldots, x_k\} \subseteq \mathbb{R}^d$ and let $\delta = \min_{i \neq j} \|x_i - x_j\|$. Then there exists a universal constant $C > 0$ such that*

$$\mathcal{E}_1^2(\mu,\nu) \geq \frac{C\delta}{k\sqrt{d}}\mathsf{TV}^2(\mu,\nu).$$

Proof. Let $\mu = \sum_{i=1}^k \mu_i \delta_{x_i}$ and $\nu = \sum_{i=1}^k \nu_i \delta_{x_i}$. Then, by [15, Theorem 1] we have

$$\mathcal{E}_1^2(\mu,\nu) = -\sum_{i,j}(\mu_i - \nu_i)(\mu_j - \nu_j)\|x_i - x_j\| \geq \frac{C\delta}{\sqrt{d}}\sum_{i=1}^k(\mu_i - \nu_i)^2 \geq \frac{C\delta\mathsf{TV}^2(\mu,\nu)}{k\sqrt{d}}$$

as required.

□

Notice that by our discussion above, the support set $\{x_1, \ldots, x_k\}$ in Theorem 1.5.5 is arbitrary and may be chosen by us. Since the scale of the supporting points $x_1, \ldots, x_k$ is statistically irrelevant, we remove this ambiguity by restricting the points to lie in the unit ball, that is, we require that $\max_i \|x_i\| \leq 1$. We see now that the comparison between $\mathcal{E}_1$ and TV *improves* as $\delta/\sqrt{d}$ grows. Given a fixed value of δ, we want to make the dimension d of

our embedding as low as possible, which means that the points $x_1, \dots, x_k$ should form a large δ-packing of the d-dimensional unit ball. Due to well known bounds on the packing number of the Euclidean ball, it follows that the best one can hope for is

$$\log(k) \asymp d \log(1/\delta).$$

Maximizing $\delta/\sqrt{d}$ subject to this constraint yields the choice $d = \Theta(\log(k))$ and $\delta = \Theta(1)$. This gives us the following corollary.

Corollary 1.5.6. *There exists a universal constant $C \in (0, \infty)$ such that for any $k \geq 1$ there exists a set of points $x_1, \dots, x_k \in \mathbb{R}^{\lceil C \log(k) \rceil}$ with $\max_i \|x_i\| \leq 1$ such that*

$$\mathcal{E}_1 \left(\sum_{i=1}^k \mu_i \delta_{x_i}, \sum_{i=1}^k \nu_i \delta_{x_i} \right) \geq \frac{\mathsf{TV}(\mu, \nu)}{C \sqrt{k} \sqrt[4]{\log(k)}}$$

for any two probability mass functions $\mu = (\mu_1, \dots, \mu_k)$ and $\nu = (\nu_1, \dots, \nu_k)$.

The question arises how the set of points $x_1, \dots, x_k$ in Corollary 1.5.6 should be constructed. One solution is to use an error correcting code (ECC), whereby we take the x_i to be the codewords of an ECC on the scaled hypercube $\frac{1}{\sqrt{d}}\{\pm 1\}^d$ for some dimension d, known as the "blocklength" in this context. An ECC is *asymptotically good* if the message length $\log(k)$ is linear in the blocklength d, that is $d \asymp \log(k)$, and if the minimum Hamming distance between any two codewords is $\Theta(d)$, which translates precisely into $\delta = \min_{i \neq j} \|x_i - x_j\| \asymp 1$. Many explicit constructions of asymptotically good error correcting codes exist, see [120] for one such example, and random codes are almost surely good [17]. Clearly the better the code is, the better the constants we obtain in Corollary 1.5.6.

Remark 8. *One interesting consequence of Corollary 1.5.6 and the preceeding discussion is the following: given a categorical feature with k possible values, the perceptron may obtain better performance by identifying each category with the codewords $x_1, \dots, x_k$ of an ECC instead of standard one-hot encoding. The idea of using ECCs instead of one-hot-encoding has been proposed before [65, 134, 175] along with data-dependent variants [50, 69, 204]. Our observations in this section might give some theoretical justification for these methods.*

Binomial Distributions

Finally, we look at the specific case of binomial distributions.

Proposition 1.5.7. *There exists a positive constant C such that for all $p, q \in [0, 1]$ and $n \in \mathbb{N}$ the inequality*

$$\mathcal{E}_1^2 \Big(\operatorname{Bin}(n, p), \operatorname{Bin}(n, q) \Big) \geq C \min \left\{ n|p - q|, \frac{n^2 (p - q)^2}{\sqrt{np + nq}}, n^2 (p - q)^2 \right\}$$

holds.

The proof of Proposition 1.5.7 is based on two technical lemmas, both of which are inspired by discussions with Yanjun Han about Poisson distributions.

Lemma 1.5.8. *Let $n \in \mathbb{N}, x \in \{0, 1, \ldots, n\}$ and $p \in [0, 1]$. Then*

$$\frac{\mathrm{d}}{\mathrm{d}p}\mathbb{P}(\mathrm{Bin}(n, p) \leq x) = -n\mathbb{P}(\mathrm{Bin}(n-1, p) = x). \tag{1.5.4}$$

Proof. Follows by direct calculation. □

Lemma 1.5.9. *Let $n \in \mathbb{N}, p \in [0, 1/2]$ and $x \in \{0, 1, 2, \ldots, n\}$ with $|x - np| \leq \sqrt{np}$. Then*

$$\mathbb{P}(\mathrm{Bin}(n, p) = x) \geq \frac{2.74 \times 10^{-5}}{1 + 2\sqrt{np}}. \tag{1.5.5}$$

Proof. Note that if $p = 0$ the claim is trivially true, so assume $p > 0$ without loss of generality. Let $\lambda = np$ and $x \in [\lambda \pm \sqrt{\lambda}]$. We break our argument into three cases.

1. Suppose $\lambda < (3 - \sqrt{5})/2$. Then $\lambda + \sqrt{\lambda} < 1$, so the only valid choice for x is 0. Plugging in, we have

$$\mathbb{P}(\mathrm{Bin}(n, p) = 0) = (1 - p)^{\lambda/p} \geq 2^{\sqrt{5}-3} \geq 0.5.$$

2. Suppose $(3 - \sqrt{5})/2 \leq \lambda < (7 - \sqrt{13})/2 \approx 1.697$. Then $\lambda + \sqrt{\lambda} < 3$ so the only valid choices for x are $0, 1$ and 2. Once again, using the inequality $1 - p \geq e^{-1.4p}$ valid for $p \in (0, 1/2)$, the following can be checked numerically:

$$\begin{aligned}
&\mathbb{P}(\mathrm{Bin}(n, p) = 0) = (1 - p)^{\lambda/p} \geq 2^{\sqrt{13}-7} && \geq 0.09 \\
&\mathbb{P}(\mathrm{Bin}(n, p) = 1) = \lambda(1 - p)^{\lambda/p-1} \geq \lambda e^{-1.4\lambda+1.4p} && \geq 0.2 \\
&\mathbb{P}(\mathrm{Bin}(n, p) = 2) = \frac{1}{2}(\lambda^2 - \lambda p)(1 - p)^{\lambda/p-2} \geq \frac{1}{2}(\lambda^2 - \frac{\lambda}{2})e^{-1.4\lambda+2.8p} && \geq 0.03.
\end{aligned}$$

3. Suppose $\lambda \geq (7 - \sqrt{13})/2$ and $x, y \in [\lambda \pm \sqrt{\lambda}] \cap \{0, 1, \ldots, n\}$ with $x \neq y$. Then

$$\begin{aligned}
\frac{\mathbb{P}(\mathrm{Bin}(n, p) = x)}{\mathbb{P}(\mathrm{Bin}(n, p) = y)} = \frac{y!(n-y)!}{x!(n-x)!}p^{x-y}(1-p)^{y-x} &= \prod_{t=\min\{x,y\}+1}^{\max\{x,y\}} \frac{1-p}{p}\frac{t}{n+1-t} \\
&\leq \left(\frac{1-p}{p}\frac{\lambda+\sqrt{\lambda}}{n-\lambda-\sqrt{\lambda}}\right)^{|x-y|} = \left(\frac{1+\frac{1}{\sqrt{\lambda}}}{1-\frac{p}{1-p}\frac{1}{\sqrt{\lambda}}}\right)^{|x-y|} \\
&\leq \left(\frac{1+\frac{1}{\sqrt{\lambda}}}{1-\frac{1}{\sqrt{\lambda}}}\right)^{|x-y|} \leq \exp\left(\frac{2|x-y|}{\sqrt{\lambda}-1}\right),
\end{aligned}$$

where we used the inequality $1 + x \leq \exp(x)$ and $p/(1 - p) \leq 1$. Since $|x - y| \leq 2\sqrt{\lambda}$ and $\lambda \geq 1.697$, we have

$$\exp\left(\frac{2|x-y|}{\sqrt{\lambda}-1}\right) \leq \exp\left(4 + \frac{4}{\sqrt{1.697}-1}\right) \leq 5473.$$

By Chebyshev's inequality we know that

$$\sum_{x\in[np\pm 2\sqrt{np}]} \mathbb{P}(\mathrm{Bin}(n,p)=x) \geq \frac{3}{4}.$$

One of the modes of the binomial distribution is always one of the two integers closest to np, see for example [121]. Moreover, in the case $\lambda \geq (7-\sqrt{13})/2$ that we are considering, the interval $[np \pm \sqrt{np}]$ always contains the two closest integers to np. As a consequence, the mode always belongs to $[np \pm \sqrt{np}]$. Since $[np \pm \sqrt{np}]$ always contains at least $\sqrt{np}$ points, and $[np \pm 2\sqrt{np}]$ contains at most $1+4\sqrt{np}$ points, we get

$$\sum_{x\in[np\pm \sqrt{np}]} \mathbb{P}(\mathrm{Bin}(n,p)=x) \geq \frac{3}{20}.$$

From here we immediately get

$$\mathbb{P}(\mathrm{Bin}(n,p)=x) \geq \frac{3}{20\times 5473}\frac{1}{1+2\sqrt{np}},$$

concluding the proof.

□

Remark 9. *If $p \in (1/2, 1]$ then one can replace $\sqrt{np}$ by $\sqrt{n(1-p)}$ and the result remains true. Moreover, up to some technicalities, an analogous statement seems to hold for all log-concave distributions in one dimension. Namely, every log-concave distribution puts constant mass on mean±deviation and the density on said range changes by at most a universal factor.*

Proof of Proposition 1.5.7. Assume without loss of generality that $p \geq q$. We further assume that $p \leq 1/2$ so that we may later apply Lemma 1.5.9. This is also without loss of generality due to the following simple reduction. Split each support element $i \in [k]$ into two, assigning exactly half of the available mass to both. This inflates the support size by a factor of two and maintains total variation distance. This is also known as the "flattening" [102, 64, 80] reduction, which we already used in Section 1.4.5 of the thesis.

From (1.2.3) and Lemma 1.5.8 we have that

$$\begin{aligned}
\mathbb{E}Z_i &= \mathcal{E}^2(\mathrm{Bin}(n,p),\mathrm{Bin}(n,q)) \\
&= \sum_{x=0}^{n}\Big(\mathbb{P}(\mathrm{Bin}(n,p)\leq x) - \mathbb{P}(\mathrm{Bin}(n,q)\leq x)\Big)^2 \\
&= n^2\sum_{x=0}^{n-1}\left(\int_q^p \mathbb{P}(\mathrm{Bin}(n-1,t)=x)\mathrm{d}t\right)^2.
\end{aligned}$$

Let us disregard the n^2 multiplier and focus on the sum. Moreover, relabel $n-1$ as n to simplify notation. Note that if $n(p+q)\leq 3$ then most mass is concentrated at 0 and we may use the simple bound

$$\sum_{x=0}^{n}\left(\int_q^p \mathbb{P}(\mathrm{Bin}(n,t)=x)\mathrm{d}t\right)^2 \geq \left(\int_q^p (1-t)^n \mathrm{d}t\right)^2 \gtrsim (p-q)^2.$$

Going forward, assume that $n(p+q) \geq 3$ holds, and note that $p \geq 1.5/n$ in this case. By Lemma 1.5.9 we obtain

$$\sum_{x=0}^{n} \left(\int_q^p \mathbb{P}(\mathrm{Bin}(n,t) = x) \mathrm{d}t \right)^2 \gtrsim \sum_{x=0}^{n} \left(\int_q^p \frac{\mathbb{1}\{x \in [nt \pm \sqrt{nt}]\}}{1+\sqrt{nt}} \mathrm{d}t \right)^2 \\ \gtrsim \frac{1}{pn} \sum_{x=0}^{n} \left(\int_q^p \mathbb{1}\{x \in [nt \pm \sqrt{nt}]\} \mathrm{d}t \right)^2.$$

Expanding the square and performing the sum over x, the above becomes

$$\asymp \frac{1}{pn} \int_{q \leq s \leq t \leq p} (ns + \sqrt{ns} - (nt - \sqrt{nt})_+)_+ \mathrm{d}s\mathrm{d}t.$$

Now, since $\alpha \mapsto \alpha - \sqrt{\alpha}$ is non-negative and increasing on $\alpha \geq 1$, we can lower bound the above integral as

$$\int_{q \leq s \leq t \leq p} (ns + \sqrt{ns} - (nt - \sqrt{nt})_+)_+ \mathrm{d}s\mathrm{d}t \geq \int_{q \leq s \leq t, 1/n \leq t \leq p} (\sqrt{nt} - n(t-s))_+ \mathrm{d}s\mathrm{d}t = (\star).$$

Suppose we further restrict the region of integration above to

$$A = \left\{ (s,t) : q \leq s \leq t, \frac{1}{n} \vee \frac{p}{2} \leq t \leq p, n(t-s) \leq \frac{1}{2}\sqrt{nt} \right\}.$$

This immediately yields the bound $(\star) \gtrsim \sqrt{np}|A|$ where $|A|$ denotes the area of A. A simple sketch of A, which is a diagonal strip in the square $[p,q]^2$, shows that $|A| \asymp (p-q)\left(\sqrt{\frac{p}{n}} \wedge (p-q)\right)$. Summarizing, we obtain the bound

$$\mathcal{E}^2(\mathrm{Bin}(n,p), \mathrm{Bin}(n,q)) \gtrsim \begin{cases} n^2(p-q)^2 & \text{if } n(p+q) \leq 3 \\ (n|p-q|) \wedge \frac{n^2(p-q)^2}{\sqrt{np}} & \text{otherwise.} \end{cases}$$

By checking the cases arising from $n|p-q| \lesssim 1$ and $np \lesssim 1$, one can verify that the above bound is equal, up to constant, to the claimed result. □

1.5.2 Density Estimation Using Half-Spaces

Suppose that we observe data $X_1, \ldots, X_n \overset{iid}{\sim} \nu$ from some unknown ν which is known to belong to a class of distributions $\mathcal{P}$. We define the minimum perceptron discrepancy estimator $\tilde{\nu}$ as

$$\tilde{\nu} \in \operatorname*{arg\,min}_{\nu' \in \mathcal{P}} \overline{d_H}(\nu', \nu_n), \tag{1.5.6}$$

where $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the empirical measure of our sample. In this section we analyse the performance of $\tilde{\nu}$ when applied to smooth distributions, Gaussian mixtures and discrete distributions. The definition of $\tilde{\nu}$ is reminiscent of the idea underpinning Generative Adversarial Networks (GANs), where the density estimate is trained in conjunction with a discriminator.

Here the discriminator is given by a one-layer network with threshold activation. In practice the discriminator would likely be parametrized by a deep neural network, thus (1.5.6) can be regarded as the simplest instantiation of this idea.

Assume for a moment that we have an inequality of the form

$$\mathsf{TV}(\mu,\nu)^\alpha \le c\,\overline{d_H}(\mu,\nu) \tag{1.5.7}$$

valid for all $\mu,\nu\in\mathcal{P}$ and some finite constants c,α that depend only on $\mathcal{P}$. Note that $\alpha\ge 1$ must necessarily hold due to the inequality $\overline{d_H}\le\mathsf{TV}$. We claim that then

$$\mathbb{E}\mathsf{TV}(\tilde\nu,\nu)\le c'(1/n)^{\frac{1}{2\alpha}} \tag{1.5.8}$$

for a finite constant c' depending only on $\mathcal{P}$. Recall that the VC-dimension of the set of half-space indicators in $\mathbb{R}^d$ is $d+1$, so that $\mathbb{E}\overline{d_H}(\nu,\nu_n)\lesssim\sqrt{d/n}$ for a universal implied constant. Our conclusion (1.5.8) follows immediately from taking the expectation in the chain of inequalities

$$\mathsf{TV}(\tilde\nu,\nu)^\alpha \overset{\text{Eq. (1.5.7)}}{\lesssim} \overline{d_H}(\tilde\nu,\nu) \overset{\text{Eq. (1.5.6)}}{\le} 2\overline{d_H}(\nu,\nu_n)$$

and applying Jensen's inequality. This brings us to the following informal result.

Theorem 1.5.10 (informal)**.** *The minimum half-space discrepancy density estimator $\tilde\nu$ defined in (1.5.6) achieves the rate $n^{-\frac{\beta}{2\beta+d+1}}$ over the class $\mathcal{P}_S(\beta,d,C)$ and the rate $(\log n)^{\frac{d+1}{2}}/\sqrt{n}$ over the class $\mathcal{P}_G(d)$.*

Sketch proof. Combine Lemmas 1.5.2 and 1.5.4 with (1.5.8). □

Both rates in Theorem 1.5.10 are close to being optimal. In the case of the smoothness class $\mathcal{P}_S(\beta,d,C)$ it is well known, see for example [107], that the minimax optimal rate of estimation is given by $n^{-\beta/(2\beta+d)}$, so our estimator $\tilde\nu$ has the same performance as the optimal one in dimension $d+1$. The minimax optimal rate over $\mathcal{P}_G(d)$ is not known, but the rate in Theorem 1.5.10 is within $\log(n)^{\Theta(d)}$ of the unknown optimal rate.

We mention that the results of Theorem 1.5.10 can be improved if we were to replace $\overline{d_H}$ by the generalized energy distance $\mathcal{E}_\gamma$ in the definition of $\tilde\nu$ in (1.5.6). In particular, for the class $\mathcal{P}_S(\beta,d,C)$ it would improve the rate to $n^{-\beta/(2\beta+d+\gamma)}$, meaning that performance improves as $\gamma\downarrow 0$. There is one caveat however, namely that some of the inequalities involved in its proof deteriorate as $\gamma\downarrow 0$. It turns out the optimal trade-off is achieved by setting $\gamma=(\log(n))^{-1}$ in which case we achieve the rate $(\log(n)/n)^{\beta/(2\beta+d)}$, which is only polylog factor off from the minimax optimal rate. Analogous statements can be made also for the Gaussian mixture class $\mathcal{P}_G(d)$.

Finally, we comment on discrete distributions. The task of estimating discrete distributions is effectively trivial, as the empirical probability mass function is minimax optimal, see for example [35, Theorem 1] for an exposition of this fact. However, as a consequence of Theorem 1.5.5 we can say the following. Suppose our observations $X_1,\dots,X_n$ are from a discrete distribution with support size $k\ge 2$. If we embed the support as a $\Omega(1)$-packing of the unit ball in $\log(k)$ dimensions, then any estimator $\tilde\nu$ that satisfies $\overline{d_H}(\tilde\nu,\nu_n)\lesssim\sqrt{d/n}$ will be a minimax optimal distribution estimator up to a $\log(k)$ factor. The point of this observation is to show that approximate minimizers of the half-space discrepancy, such as those found by first order methods applied to GANs, are also good estimators.

1.5.3 Two Sample Testing Using Half-Spaces

We saw in the previous section how the perceptron discrepancy can be used to design near-optimal density estimators for smooth and Gaussian mixture distributions. In this section we apply it to a rather different problem: that of two-sample testing for smooth and discrete distributions.

Smooth Distributions

In a recent paper [161], the following test statistic for two-sample testing was proposed:

$$T_{d,k}(p,q) = \max_{(w,b)\in\mathbb{S}^{d-1}\times[0,\infty)} \left| \mathbb{E}_{X\sim p}\left[\mathrm{Relu}(\langle w, X\rangle - b)^k\right] - \mathbb{E}_{Y\sim q}\left[\mathrm{Relu}(\langle w, Y\rangle - b)^k\right] \right|,$$

where $\mathrm{Relu}(x) = \max\{x, 0\}$ as usual. Notice that clearly $T_{d,0} \equiv \overline{d_H}$. They propose to reject the null hypothesis that $p = q$ whenever

$$T_{d,k}(p_n, q_n) \geq t_n, \tag{1.5.9}$$

where $p_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $q_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ are empirical measures of two i.i.d. samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, and the threshold satisfies both $t_n = o(1)$ and $t_n = \omega(1/\sqrt{n})$. One of their main technical results [161, Theorem 6] shows that the test (1.5.9) returns the correct hypothesis with probability $1 - o(1)$ asymptotically as $n \to \infty$ for any qualifying sequence $\{t_n\}_{n\geq 1}$ and fixed p, q. However, this result leaves open questions about the sample complexity of their test, and in particular, whether it is able to achieve known minimax rates.

We show for $k = 0$ that (1.5.9) is *not* able to obtain the minimax optimal sample complexity over the smoothness class $\mathcal{P}_{\mathsf{H}}(\beta, d, C)$ introduced in Section 1.2.3. It is well known (see for example [10, 139]) that two-sample testing over $\mathcal{P}_{\mathsf{H}}(\beta, d, C)$ is solvable with probability $1 - o(1)$ if and only if

$$n = \omega(\epsilon^{-\frac{2\beta+d/2}{\beta}}), \tag{1.5.10}$$

in which case a variant of the χ^2-test, pioneered by Ingster, works. It turns out that our construction showing the tightness of our comparison between $\overline{d_H}$ and TV in Lemma 1.5.3, also shows that (1.5.9) cannot achieve the optimal sample complexity (1.5.10).

Proposition 1.5.11 ([75, Proposition 10])**.** *For all $d, \beta > 0$, there exists constants c, c' such that for all $\epsilon > 0$, there exists probability density functions p, q supported on the d-dimensional unit ball such that*

1. $\|p\|_{\beta,2}, \|q\|_{\beta,2} \leq c$,
2. $\|p - q\|_1 \asymp \|p - q\|_2 \asymp \epsilon$, *and*
3. *the expected test statistic satisfies*

$$\mathbb{E}[T_{d,0}(p_n, q_n)] \leq \frac{c'}{\sqrt{n}}$$

for any $n \leq (\log \frac{1}{\epsilon})^{-d}\epsilon^{-\frac{2\beta+d+1}{\beta}}$.

In other words, consistent testing using the statistic $T_{d,0}$ is impossible with $n = \tilde{o}(\epsilon^{-\frac{2\beta+d+1}{\beta}})$ samples, which is a far cry from the optimal sample complexity (1.5.10).

Discrete Distributions

In the previous section we showed that thresholding $\overline{d_H}$ does not result in a minimax optimal two-sample testing procedure for smooth distributions. In contrast, in this section we find that the energy distance $\mathcal{E}_1$ can be applied to the discrete two-sample testing problem to obtain optimal performance.

Let p, q be two probability mass functions on the alphabet $[k] = \{1, 2, \dots, k\}$ and suppose that $A_1, \dots, A_{2n}$ and $B_1, \dots, B_{2n}$ are two i.i.d. samples from p, q respectively. Let $X_i = \sum_{j=1}^n \mathbb{1}\{A_j = i\}$, $X_i' = \sum_{j=1}^n \mathbb{1}\{X_{n+j} = i\}$ and define Y_i, Y_i' in terms of the B_j analogously. Our goal is to present a novel analysis of the statistic

$$T := \sum_{i=1}^k \Big[|X_i - Y_i| + |X_i' - Y_i'| - |X_i - X_i'| - |Y_i - Y_i'| \Big],$$

which was introduced by [62]. The motivation of the authors for introducing the statistic is a bit mysterious, and seems to have been the product of educated guesswork. However, in light of Section 1.2.4 we see that $\mathbb{E}T$ is simply the sum of energy distances between two binomials, each term corresponding to one support element. To make this more explicit, writing $T = \sum_{i=1}^k Z_i$, we have

$$\mathbb{E}Z_i = \mathcal{E}^2(\mathrm{Bin}(n, p_i), \mathrm{Bin}(n, q_i)).$$

The concentration of T is straightforward, and the bulk of the technical difficulty lies in lower bounding the expectation under the alternative hypothesis (the mean is clearly zero under the null). Indeed, by McDiarmid's inequality there exists a universal constant $C > 0$ such that

$$\mathbb{P}(|T - \mathbb{E}T| \geq t) \leq 2\exp(-Ct^2 n)$$

for any $t \geq 0$. In other words, T is $\mathcal{O}(1/n)$-sub-Gaussian. To deal with the mean, [62] proves the following.

Proposition 1.5.12 ([62, Claim 3.4])**.** *There exists a universal constant $C > 0$ such that for any $i \in [k]$,*

$$\mathbb{E}Z_i \geq C \min\left\{ n|p_i - q_i|, \frac{n^2(p_i - q_i)^2}{\sqrt{n(p_i + q_i)}}, n^2(p_i - q_i)^2 \right\}. \tag{1.5.11}$$

The original proof of the key Proposition 1.5.12 proceeds by Poissonizing the samples, and using some clever tricks, that may appear ad-hoc. More recently, a short (12 page) note [37] was published which gave an alternative, more direct/principled analysis of $\mathbb{E}T$ using the fact that

$$\mathbb{E}|W| = \frac{2}{\pi} \int_0^\infty \frac{1 - \mathbb{E}\cos(itW)}{t^2} \mathrm{d}t$$

for any integrable random variable W, which is attributed to Zolotarev [166], see also [190, Chapter 9] for historical details. We note that using the Zolotarev identity to study $\mathbb{E}T$ is essentially the same as using the Fourier-form (Proposition 1.2.1) of the energy distance, although this connection isn't made in [37]. However, we can recognize Proposition 1.5.12 as

just re-stating our separation result Proposition 1.5.7, whose proof provides a third, in our opinion further simplified proof of the corresponding results of [62].

For completeness, we mention the remaining steps needed to derive the minimax sample complexity of two-sample testing. Partition the indices $[k]$ into three sets $S_1 \sqcup S_2 \sqcup S_3 = [k]$ according to which term on the RHS of (1.5.11) attains the minimum, breaking ties arbitrarily. Since $\mathsf{TV}(p, q) \geq \epsilon$ under the alternative hypothesis, we have that

$$\max_{j=1,2,3} \sum_{i \in S_j} |p_i - q_i| \geq 2\epsilon/3. \tag{1.5.12}$$

Depending on whether $j = 1, 2$ or 3 attains the maximum in (1.5.12), a straightforward calculation using Hölder's inequality shows that

$$\mathbb{E}T = \sum_{i=1}^{n} \mathbb{E}Z_i \geq Cn \min\left\{\epsilon, \epsilon^2 \frac{n}{k}, \epsilon^2 \sqrt{\frac{n}{k}}\right\},$$

for a universal constant $C > 0$. Combining the above with the sub-Gaussianity of T, which says that $T = \mathbb{E}T \pm \mathcal{O}(\sqrt{\log(1/\delta)/n})$ with probability at least $1 - \delta$, we see that to perform two-sample testing with type-I + type-II error bounded by δ it is sufficient to take

$$n \gtrsim \frac{\log(1/\delta)}{\epsilon^2} + \frac{\sqrt{k \log(1/\delta)}}{\epsilon^2} + \frac{k^{2/3} \log^{1/3}(1/\delta)}{\epsilon^{4/3}}$$

observations, which can be shown to be minimax optimal.

1.5.4 Separating Distributions by Arbitrary Sets

In Section 1.5.1 we looked at separating distributions by the best possible half-space (using $\overline{d_H}$), and random half-spaces (using the energy distance). In this section we take a more flexible approach, and don't restrict ourselves to half-spaces. Instead, given samples, we carefully construct sets where the two distributions are guaranteed to differ with high probability.

Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are i.i.d. from P_X and P_Y respectively. Our goal is to construct a set $\widehat{S}$ based on these observations such that the separating power $\mathrm{sep}(\widehat{S})$ is "large" and the size $\tau(\widehat{S})$ is "small", where

$$\mathrm{sep}(\widehat{S}) = P_X(\widehat{S}) - P_Y(\widehat{S})$$

and

$$\tau(\widehat{S}) = \min\left\{P_X(\widehat{S})P_X(\widehat{S}^{\mathsf{c}}), P_Y(\widehat{S})P_Y(\widehat{S}^{\mathsf{c}})\right\}.$$

We shall see the importance of sep and τ later in Section 1.5.5, where we apply our results downstream to goodness-of-fit, two-sample, and likelihood-free testing.

Gaussian Sequence Model

Our approach to the Gaussian sequence model is quite natural: $\widehat{S}$ is simply a superlevel set of a truncated version of the likelihood ratio between estimates of P_X and P_Y. Indeed, suppose we

have two samples X, Y of size n from $\otimes_{j=1}^{\infty}\mathcal{N}(\theta_j^X, 1) =: \mu_{\theta^X}$ and μ_{θ^Y} respectively, where θ^X, θ^Y have Sobolev norm $\|\theta\|_s^2 =: \sum_j \theta_j^2 j^{2s}$ bounded by a constant and satisfy $\mathsf{TV}(\mu_{\theta^X}, \mu_{\theta^Y}) \geq \epsilon > 0$. We use $\widehat{\theta}^X$ and $\widehat{\theta}^Y$ to denote the empirical mean vector of the samples X and Y respectively.

The separating set is constructed as follows:

$$\widehat{S} = \{Z \in \mathbb{R}^{\mathbb{N}} : T(Z) \geq 0\},$$

where $T(Z) = 2\sum_{j=1}^{J}(\widehat{\theta}_j^X - \widehat{\theta}_j^Y)(Z_j - (\widehat{\theta}_j^X + \widehat{\theta}_j^Y)/2)$ for some $J \in \mathbb{N}$ to be specified. This is simply a truncated version of the likelihood-ratio test between $\mu_{\widehat{\theta}^X}$ and $\mu_{\widehat{\theta}^Y}$, where we set all but the first J coordinates of $\widehat{\theta}^X$ and $\widehat{\theta}^Y$ to zero. The performance of the separating set is summarized in the next proposition.

Proposition 1.5.13 ([74, Proposition 14]). *There exists universal constants c, c' such that when $J = \lfloor c\epsilon^{-1/s} \rfloor$ the inequality*

$$\mathbb{P}\left(\mu_{\theta^X}(\widehat{S}) - \mu_{\theta^Y}(\widehat{S}) \geq c'\left(\sqrt{n\epsilon^{1/s}} \wedge \frac{1}{\epsilon}\right)\epsilon^2\right) \geq 1 - \delta$$

holds, provided $n \gtrsim \frac{1}{c'} n_{\mathsf{TS}}(\epsilon, \delta, \mathcal{P}_{\mathsf{G}})$.

The minimax sample complexity $n_{\mathsf{TS}}(\epsilon, \delta, \mathcal{P}_{\mathsf{G}})$ is recorded in Table 1.2. The main tool in the proof of Proposition 1.5.13 is Gaussian Lipschitz concentration.

Discrete Distributions

Suppose now that P_X, P_Y are supported on $[k] := \{1, 2, \ldots, k\}$ with p.m.f.s p_X, p_Y respectively. One natural and simple approach is to take $\widehat{S}_{1/2}$ defined as

$$i \in \widehat{S}_{1/2} \iff \widehat{p}_X(i) + \frac{1}{2n}\Omega_i > \widehat{p}_Y(i),$$

where $\Omega_1, \ldots, \Omega_k$ are i.i.d. Rademacher variables independent of the data and $\widehat{p}_X, \widehat{p}_Y$ are the empirical p.m.f.s. In other words, given a support element $i \in [k]$, classify it as X if its empirical frequency is larger in the X-sample and to flip a coin in case of a tie. This "classifier" can also be thought as a minimizer of the empirical misclassification error on our training set.

The bound $\tau(\widehat{S}_{1/2}) \leq 1/4$ holds trivially. For the separation $\mathrm{sep}(\widehat{S}_{1/2})$ we must work a bit harder. The following is the key technical result for this step.

Lemma 1.5.14 ([74, Lemma 21]). *Let $\mu \geq \lambda \geq 0$ and $X \sim \mathrm{Poi}(\mu), Y \sim \mathrm{Poi}(\lambda)$. Then*

$$\mathbb{P}(X > Y) + \frac{1}{2}\mathbb{P}(X = Y) - \frac{1}{2} \geq c\left(\frac{\mu - \lambda}{\sqrt{\lambda + 1}} \wedge 1\right)$$

holds, where $c > 0$ is a universal constant.

Assuming, without loss of generality, that our samples are Poissonized, we can proceed:

$$\begin{aligned}\mathbb{E}\,\mathrm{sep}(\widehat{S}_{1/2}) &= \sum_{i=1}^{k} \big(p_X(i) - p_Y(i)\big)\mathbb{P}\left(i \in \widehat{S}_{1/2}\right) \\ &= \sum_{i=1}^{k} \big(p_X(i) - p_Y(i)\big)\left(\mathbb{P}\left(i \in \widehat{S}_{1/2}\right) - \frac{1}{2}\right) \\ &\overset{\text{Lem. 1.5.14}}{\geq} c\sum_{i=1}^{k} \big|p_X(i) - p_Y(i)\big|\left(\frac{n|p_X(i) - p_Y(i)|}{\sqrt{n(p_X(i) \wedge p_Y(i)) + 1}} \wedge 1\right),\end{aligned}$$

where the last step uses that

$$\mathbb{P}\left(i \in \widehat{S}_{1/2}\right) = \mathbb{P}\left(\mathrm{Poi}\left(np_X(i)\right) > \mathrm{Poi}\left(np_Y(i)\right)\right) + \frac{1}{2}\mathbb{P}\left(\mathrm{Poi}\left(np_X(i)\right) = \mathrm{Poi}\left(np_Y(i)\right)\right).$$

In the above display we use Poissonized samples so our X and Y sample may have a different number of total observations, although both are guaranteed to be $n + \mathcal{O}(\sqrt{n})$ with high probability. This is in contrast with how we defined $\widehat{S}_{1/2}$ above, where the size of the two samples is fixed and equal, but the difference between the two is negligible due to the concentration of the Poisson distribution.

Under the assumption that $\|p_X - p_Y\|_1 \geq \epsilon$, we can analyse the three cases that may arise for each term in the sum above, and we arrive at the bound

$$\mathbb{E}\,\mathrm{sep}(\widehat{S}_{1/2}) \gtrsim \epsilon^2\left(\frac{1}{\epsilon} \wedge \sqrt{\frac{n}{k}} \wedge \frac{n}{k}\right) =: \mathrm{sep}^*.$$

To summarize, so far we have the bound $\tau(\widehat{S}_{1/2}) \leq 1/4$ with probability one, and a lower bound on $\mathbb{E}\,\mathrm{sep}(\widehat{S}_{1/2})$. To ensure that the latter bound holds not only in expectation but with probability $1 - \delta$, we must analyse the concentration properties of $\mathrm{sep}(\widehat{S}_{1/2})$. To achieve the optimal sample complexity, we rely on the exact characterization of the sub-Gaussian variance proxy of Bernoulli random variables.

Lemma 1.5.15 ([34, Theorem 2.1])**.** *Let $\sigma^2_{\mathsf{opt}}(\mu)$ be the optimal (smallest) sub-Gaussian variance proxy of the* $\mathrm{Ber}(\mu)$ *distribution. Then*

$$\sigma^2_{\mathsf{opt}}(\mu) = \frac{\frac{1}{2} - \mu}{\log\left(\frac{1}{\mu} - 1\right)},$$

where the values for $\mu \in \{0, \frac{1}{2}, 1\}$ should be understood as the limit of the above expression, resulting in $\sigma^2_{\mathsf{opt}} = 0, \frac{1}{4}, 0$ respectively.

Still assuming that our observations are Poissonized, by standard tail bounds we know that for any $i \in [k]$

$$\mathsf{p}_i := \mathbb{P}(i \in \widehat{S}_{1/2}) \wedge \mathbb{P}(i \notin \widehat{S}_{1/2}) \leq 2\exp\left(-n\frac{(p_X(i) - p_Y(i))^2}{p_X(i) + p_Y(i)}\right) \tag{1.5.13}$$

holds. Using Lemma 1.5.15 we can bound the sub-Gaussian variance proxy as

$$\begin{aligned}\sigma^2_{\mathsf{opt}}\left(\mathrm{sep}(\widehat{S}_{1/2})\right) &= \sigma^2_{\mathsf{opt}}\left(\sum_{i=1}^{n}(p_X(i)-p_Y(i))\mathbb{1}\{i\in\widehat{S}_{1/2}\}\right)\\ &\overset{\text{Lem. 1.5.15}}{=} \sum_{i=1}^{k}(p_X(i)-p_Y(i))^2\times\frac{\frac{1}{2}-\mathsf{p}_i}{\log\left(\frac{1}{\mathsf{p}_i}-1\right)}\\ &\overset{(1.5.13)}{\lesssim} \sum_{i=1}^{k}(p_X(i)-p_Y(i))^2\wedge\frac{p_X(i)+p_Y(i)}{n} = \mathcal{O}\left(\frac{1}{n}\right).\end{aligned}$$

Remarkably, the inverse logarithmic dependence of σ^2_{opt} for small Bernoulli parameters is crucial in obtaining the above result. Putting our bounds together, we obtain

$$\mathbb{P}\left(\tau(\widehat{S}_{1/2})\le 1/4 \text{ and } \mathrm{sep}(\widehat{S}_{1/2})\gtrsim \mathrm{sep}^*\right)\ge 1-\delta, \tag{1.5.14}$$

provided that $\mathrm{sep}^*\gtrsim\sqrt{\log(1/\delta)}\sigma_{\mathsf{opt}}(\mathrm{sep}(\widehat{S}_{1/2}))$. The latter condition rearranges to

$$n\gtrsim\frac{\log(1/\delta)}{\epsilon^2}+\frac{\sqrt{k\log(1/\delta)}}{\epsilon^2}+\frac{k^{2/3}\log^{1/3}(1/\delta)}{\epsilon^{4/3}},$$

which may be recognized as the sample complexity of two-sample testing over $\mathcal{P}_{\mathsf{D}}$.

In (1.5.14) we obtained a surprisingly simple and poweful result, however we weren't able to achieve non-trivial control over $\tau(\widehat{S})$. It is in fact possible to do better, if we assume that the distributions P_X, P_Y are bounded, that is, if $P_X, P_Y\in\mathcal{P}_{\mathsf{Db}}$.

The first improvement that one can make to $\widehat{S}_{1/2}$ is to consider the two sets $\widehat{S}_{\lessgtr}$ defined by

$$i\in\widehat{S}_{\lessgtr}\iff\widehat{p}_X(i)\lessgtr\widehat{p}_Y(i).$$

Notice that for i to be added to either set, we must observe at least one occurence of i in either the X or Y sample. This gives us an improved bound on τ of the form

$$\tau(\widehat{S}_{\lessgtr})\le\frac{1}{4}\wedge\left(n\max_{i\in\widehat{S}_{\lessgtr}}\left\{p_X(i)+q_X(i)\right\}\right),$$

which holds with high probability for Poissonized samples. If we assume that $p_X, p_Y\in\mathcal{P}_{\mathsf{Db}}$ then we obtain $\tau(\widehat{S}_{\lessgtr})\lesssim 1\wedge(n/k)$, which is exactly what we need to fill the gaps left by $\widehat{S}_{1/2}$ for this class. However, there is one caveat: the expected separation of either set is no longer guaranteed, and in fact it may even be negative!

Proposition 1.5.16 ([74, Proposition 8])**.** *Consider the distributions p, q on $[3k]$ with $p_i=\mathbb{1}\{i\le k\}/(2k)+\mathbb{1}\{i>k\}/(4k)$ and $q_i=\mathbb{1}\{i\le k\}/k$. Then, for $n\le 0.6k$,*

$$\mathbb{E}\,\mathrm{sep}(\widehat{S}_{>})<0.$$

Sketch Proof. The intuition for the construction is as follows: half the mass of p is placed on the support elements $[k]$, each of which incurs a separation loss of $-1/(2k)$. The other half of its mass is put on the support elements $(k, 3k]$, each of which incurs a separation $+1/(4k)$. Therefore, the total expected separation of $\widehat{S}_{>}$ is roughly $\frac{1}{2}\times(\frac{-1}{2k}+\frac{1}{4k})=-\frac{1}{8k}$. □

What saves us is the fact that at least one of the two sets $\widehat{S}_>, \widehat{S}_<$ has good expected separation, since $\mathbb{E}\,\mathrm{sep}(\widehat{S}_>) + \mathbb{E}\,\mathrm{sep}(\widehat{S}_<) = 2\mathbb{E}\,\mathrm{sep}(\widehat{S}_{1/2})$. In light of this, the strategy is clear: hold out a linear fraction of the training samples to decide between $\widehat{S}_>$ versus $\widehat{S}_<$ by comparing their empirical separations. By Bernstein's inequality it is then straightforward to prove that we identify the correct separating set with high probability, which results in the following.

Proposition 1.5.17 ([74, Corollary 10]). *Suppose that $P_X, P_Y \in \mathcal{P}_{\mathsf{Db}}$ with $\mathsf{TV}(p,q) \geq \epsilon$. There exists a universal constant $c > 0$ such that using the samples X, Y we can find a set $\widehat{S} \subseteq [k]$ which, with probability $1-\delta$, satisfies*

$$\left|\mathrm{sep}(\widehat{S})\right| \geq c\epsilon^2 \left(\frac{1}{\epsilon} \wedge \sqrt{\frac{n}{k}} \wedge \frac{n}{k}\right) \qquad \text{and} \qquad \tau(\widehat{S}) \leq \frac{1}{c}\left(1 \wedge \frac{n}{k}\right),$$

provided $n \geq \frac{1}{c} n_{\mathsf{GoF}}(\epsilon, \delta, \mathcal{P}_{\mathsf{Db}})$.

The above discussion covers bounded discrete distributions. However, it is of no help for the family of all discrete distributions $\mathcal{P}_{\mathsf{D}}$, as we no longer have a bound on $\max_{i \in \widehat{S}_\leqslant}(p_X(i)+q_X(i))$, so our control of τ is still trivial. The key idea here is to use yet another part of our training samples to partition the support $[k]$ into $\mathcal{O}(\log(k))$ subsets on each of which either p or q is approximately uniform. Due to its technical nature, we defer all details to Chapter 4 of the thesis.

Smooth Distributions

Analogously to Section 1.4, our results for the smooth density class $\mathcal{P}_{\mathsf{H}}$ follow by reduction to the bounded discrete $\mathcal{P}_{\mathsf{Db}}$ case. The approximation result Lemma 1.4.2 is what makes this reduction possible. Effectively, it lets us prove a result analogous to Proposition 1.5.17 except with k replaced by $\epsilon^{1/\beta}$ throughout.

1.5.5 LFHT in the Small Error Regime

In this section we look at the idea of classifier-accuracy testing, which is a popular approach amongst practitioners of LFI [91]. We derive theoretical guarantees for this class of testing algorithms, using the classifiers that we have constructed in Section 1.5.4.

Introduction to Classifier-Accuracy Testing

Given two datasets $A_1, \ldots, A_a \overset{iid}{\sim} P_A$ and $B_1, \ldots, B_b \overset{iid}{\sim} P_B$ taking values in some space $\mathcal{X}$, and a classifier $\mathcal{C} : \mathcal{X} \to \{0,1\}$, we define the "classifier-accuracy statistic"

$$T_S(A,B) := \frac{1}{a}\sum_{i=1}^{a} \mathbb{1}\{A_i \in S\} - \frac{1}{b}\sum_{j=1}^{b} \mathbb{1}\{B_j \in S\}, \tag{1.5.15}$$

where we abbreviate $A = (A_1, \ldots, A_a), B = (B_1, \ldots, B_b)$ and where we identify the classifier $\mathcal{C}$ with the set $S = \mathcal{C}^{-1}(\{1\})$, so that $\mathcal{C}(x) = \mathbb{1}\{x \in S\}$.

Assuming that the classifier $\mathcal{C}$ aims to assign 1 to data from P_A and 0 to data from P_B, it follows that $T_S(A, B) + 1$ is simply equal to the sum of the fraction of correctly classified instances. This property lends the name "classifier-accuracy" statistic. We say that a testing procedure is a classifier-accuracy test if its output is obtained by thresholding $|T_S|$ for some classifier on independent test data.

Notice that $\mathbb{E}T_S(A, B) = 0$ if $P_A = P_B$ by design, irrespective of the classifier $\mathcal{C}$. On the other hand, if $P_A \neq P_B$ and $\mathcal{C}$ is good at distinguishing the two distributions, then one expects that $|T_S(A, B)|$ is "large". Therefore, depending on the power of $\mathcal{C}$ under the alternative, thresholding $|T_S|$ should be a reasonable test of the null hypothesis $P_A = P_B$ against the alternative $P_A \neq P_B$.

Classifier-accuracy testing is a popular method used by practitioners in likelihood-free inference. One simply trains a classifier on simulated data that distinguishes P_X from P_Y and then applies the trained classifier to the experimental data to see which way the classifier leans. More precisely, suppose we have i.i.d. samples $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$ and $Z_1, \ldots, Z_m$ from unknown P_X, P_Y, P_Z respectively. Our goal is to test the null hypothesis $P_X = P_Z$ versus $P_Y = P_Z$, and we are guaranteed that P_X and P_Y are separated, and in particular not equal to each other. Suppose that we train a classifier $\mathcal{C} = \mathbb{1}_S$ on the first half of data $X_1, \ldots, X_{n/2}, Y_1, \ldots, Y_{n/2}$. Then, using the remaining data we compute the statistic $|T_S(\{Z_i\}_{i=1}^m, \{X_{n/2+i}\}_{i=1}^{n/2})|$ and reject the null for large values of this statistic.

Remark 10. *In the specific case of testing over the class of all discrete distributions $\mathcal{P}_\mathsf{D}$ we also need to use $Z_1, \ldots, Z_{m/2}$ in our construction of S whenever $m \geq n$. This can be regarded as a sort of unsupervised step in the classifier training procedure, since we do not know whether $P_Z = P_X$ or $P_Z = P_Y$. It is open whether this is information theoretically necessary, or whether one can train purely on the X, Y samples and still attain minimax optimal performance. Note that we didn't cover this in Section 1.5.4, but the details can be found in Chapter 4.*

The same idea can also be applied to two-sample testing. Recall that in the setting of two-sample testing we have two i.i.d. samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ from P_X, P_Y respectively and the goal is to test the null $P_X = P_Y$ against the alternative that $P_X \neq P_Y$. First, split the data into two batches and train a classifier $\mathcal{C} = \mathbb{1}_S$ on $X_1, \ldots, X_{n/2}$ and $Y_1, \ldots, Y_{n/2}$. Then, on the second half, threshold the classifier-accuracy statistic, i.e. reject the null hypothesis if $|T_S(\{X_{n/2+i}\}_{i=1}^{n/2}, \{Y_{n/2+i}\}_{i=1}^{n/2})|$ is large.

Finally, since goodness-of-fit testing is a special case of two-sample testing where we have an unlimited number of observations from one of the two distributions, the algorithm for goodness-of-fit testing using the classifier accuracy statistic follows from the paragraph above.

The Key Lemma

Suppose for a moment that a classifier $\mathcal{C} = \mathbb{1}_S$ has already been found. The following lemma gives an upper bound on the number of independent test observations required for the second step, that is, accepting/rejecting based on the magnitude of $|T_S|$, to correctly identify the true hypothesis with error probability at most $\delta \in (0, 1)$.

Lemma 1.5.18. *Let P, Q be two unknown distributions. Consider the hypothesis testing problem $H_0 : P = Q$ versus an arbitrary alternative H_1. Suppose that the learner has constructed a "separating set" S such that*

$$|\operatorname{sep}(S)| = |P(S) - Q(S)| \geq \underline{\operatorname{sep}} \qquad \text{for every } (P, Q) \in H_1,$$

and

$$\tau(S) = (P(S)(1 - P(S)) \wedge (Q(S)(1 - Q(S))) \leq \overline{\tau} \qquad \text{for every } (P, Q) \in H_0 \cup H_1.$$

Then, using only the knowledge of $\overline{\tau}$ and $\underline{\operatorname{sep}}$, the classifier-accuracy test (1.5.15) with n test samples from both P and Q and an appropriate threshold achieves type-I and type-II errors at most δ, provided that

$$n \geq c \frac{\log(1/\delta)}{\underline{\operatorname{sep}}} \left(1 + \frac{\overline{\tau}}{\underline{\operatorname{sep}}}\right)$$

for a large enough universal constant $c > 0$.

Note that Lemma 1.5.18 would follow immediately by Bernstein's inequality if we were to replace $\tau(S)$ by $P(S)(1 - P(S)) + Q(S)(1 - Q(S))$. This is because clearly

$$\operatorname{var}\left(T_S(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \middle| S\right) = \frac{1}{n}\Big\{P(S)(1 - P(S)) + Q(S)(1 - Q(S))\Big\}.$$

However, the savings we get by using $\tau(S)$ as defined in Lemma 1.5.18 is crucial in obtaining optimal sample complexity bounds. We are able to derive this stronger result due to the following simple observation: when the minimum of $P(S)(1-P(S))$ and $Q(S)(1-Q(S))$ is not of the same order as the sum of the two quantities, then $\operatorname{sep}(S)$ must be large, thereby making the testing problem easier. With a good choice of the classifier, we can use Lemma 1.5.18 to derive the sample complexity results in Tables 1.2 and 1.3. The following three rules lay out the interpretation of the table.

(i) Unmarked entries denote optimal results achievable by a classifier-accuracy test.
(ii) Entries marked with (OPT) denote optimal results that are not known to be achievable by any classifier-accuracy test.
(iii) Entries marked with (CAT) denote the best known result using a classifier-accuracy test.

To derive the rsults of the table, we can simply combine our classifier constructions in Section 1.5.4 with Lemma 1.5.18. More concretely, using (1.5.14) we get that the CAT using $\widehat{S}_{1/2}$ is minimax optimal for the following cases.

(i) GoF in $\mathcal{P}_{\mathsf{Db}}$ and $\mathcal{P}_{\mathsf{D}}$ as long as $k = \mathcal{O}(\log(1/\delta)/\epsilon^4)$;
(ii) TS in $\mathcal{P}_{\mathsf{Db}}$ as long as $k = \mathcal{O}(\log(1/\delta)/\epsilon^4)$, and in $\mathcal{P}_{\mathsf{D}}$ for all (k, ϵ, δ);
(iii) LFHT in $\mathcal{P}_{\mathsf{Db}}$ as long as $k = \mathcal{O}(\log(1/\delta)/\epsilon^4)$, and in $\mathcal{P}_{\mathsf{D}}$ as long as $n \geq m$.

Most notably, we see that the trivial classifier $\widehat{S}_{1/2}$ is enough to recover the high-probability minimax sample complexity of two-sample testing, which was resolved only recently in [62]. By applying Proposition 1.5.17 with Lemma 1.5.18 we can further resolve all cases for the bounded discrete class $\mathcal{P}_{\mathsf{Db}}$. In order to obtain the best results for $\mathcal{P}_{\mathsf{D}}$ and complete the table, we refer the reader to Chapter 4 or the original source [74].

	n_{GoF}	n_{TS}	$\mathcal{R}_{\mathsf{LF}}$
$\mathcal{P}_{\mathsf{Db}}(k)$	$\frac{\sqrt{k\log(1/\delta)}}{\epsilon^2} + \frac{\log(1/\delta)}{\epsilon^2}$	n_{GoF}	$m \geq \frac{\log(1/\delta)}{\epsilon^2}$ and $n \geq n_{\mathsf{GoF}}$ and $nm \geq n_{\mathsf{GoF}}^2$
$\mathcal{P}_{\mathsf{H}}(\beta, d)$	$\frac{\sqrt{\log(1/\delta)}}{\epsilon^{(2\beta+d/2)/\beta}} + \frac{\log(1/\delta)}{\epsilon^2}$	n_{GoF}	$m \geq \frac{\log(1/\delta)}{\epsilon^2}$ and $n \geq n_{\mathsf{GoF}}$ and $nm \geq n_{\mathsf{GoF}}^2$
$\mathcal{P}_{\mathsf{G}}(s)$	$\frac{\sqrt{\log(1/\delta)}}{\epsilon^{(2s+1/2)/s}} + \frac{\log(1/\delta)}{\epsilon^2}$	n_{GoF}	$m \geq \frac{\log(1/\delta)}{\epsilon^2}$ and $n \geq n_{\mathsf{GoF}}$ and $nm \geq n_{\mathsf{GoF}}^2$

Table 1.2: Minimax sample complexity of testing (up to constant factors) over $\mathcal{P}_{\mathsf{H}}, \mathcal{P}_{\mathsf{G}}, \mathcal{P}_{\mathsf{Db}}$.

	$n_{\mathsf{GoF}}(\mathcal{P}_{\mathsf{D}})$	$n_{\mathsf{TS}}(\mathcal{P}_{\mathsf{D}})$	$\mathcal{R}_{\mathsf{LF}}(\mathcal{P}_{\mathsf{D}})$	
$k \geq \frac{\log\left(\frac{1}{\delta}\right)}{\epsilon^4}$	(OPT) $n_{\mathsf{GoF}}(\mathcal{P}_{\mathsf{Db}})$	$\left(\frac{k^2\log\left(\frac{1}{\delta}\right)}{\epsilon^4}\right)^{\frac{1}{3}}$	$n \geq m$	$m \geq \frac{\log(1/\delta)}{\epsilon^2}$ and $m\min\{n^2/k, n\} \geq n_{\mathsf{GoF}}^2$
			$m > n$	(OPT) $mn^2 \geq kn_{\mathsf{GoF}}^2$ and $n \geq n_{\mathsf{GoF}}$
	(CAT) $n_{\mathsf{GoF}}\left(\frac{\epsilon}{\log(k)}, \frac{\delta}{k}, \mathcal{P}_{\mathsf{Db}}\right)$			(CAT) $\frac{mn^2}{\log\left(\frac{k}{\delta}\right)} \geq kn_{\mathsf{GoF}}^2\left(\frac{\epsilon}{\log(k)}, \frac{\delta}{k}\right)$ and $n \geq n_{\mathsf{GoF}}(\frac{\epsilon}{\log(k)}, \frac{\delta}{k})$
$k < \frac{\log\left(\frac{1}{\delta}\right)}{\epsilon^4}$	$n_{\mathsf{GoF}}(\mathcal{P}_{\mathsf{Db}})$	$n_{\mathsf{GoF}}(\mathcal{P}_{\mathsf{Db}})$	$m \geq \frac{\log(1/\delta)}{\epsilon^2}$ and $n \geq n_{\mathsf{GoF}}$ and $nm \geq n_{\mathsf{GoF}}^2$	

Table 1.3: Minimax sample complexity of testing (up to constant factors) over $\mathcal{P}_{\mathsf{D}}$.

1.6 Kernel-Based Tests for LFHT and the Empirical Trade-Off

This section is based on [76] which is joint work with Tianze Jiang, Yury Polyanskiy and Rui Sun, and was published at NeurIPS '23, and is included in Chapter 3 in full. While the previous two sections characterized the region $\mathcal{R}_{\mathsf{LF}}$ in great generality, the minimax optimal tests we exhibited were hardly usable in realistic scenarios. For example, the class $\mathcal{P}_{\mathsf{H}}$ of smooth distributions requires discretizing the observations over a grid that is of exponential size in the dimension. This prompted us to consider a practically more realistic, kernel-based procedure.

1.6.1 A Generalization of LFHT

A prominent application of likelihood-free inference lies in the field of particle physics. Scientists run sophisticated experiments in the hope of confirming the existence of a hypothesized particle or phenomenon. Often said phenomenon can be predicted from theory, and thus can be simulated, as was the case for the Higgs boson whose existence was verified after nearly 50 years at the Large Hadron Collider (LHC) [40, 5]. In such experiments, the goal is to prove that the rate at which a special particle (such as the Higgs boson) is born is greater than 0, or more ambitiously, is lower bounded by a positive quantity that is within the range of theoretical predictions. The generalization of LFHT that we introduce next, which we call mixed LFHT (mLFHT), was inspired by this rate-detection problem.

Suppose we have n simulations from the background distribution P_X and the signal distribution P_Y. Further, we also have m datapoints from $P_Z = (1-\nu)P_X + \nu P_Y$, so the observed data is a mixture between the background and signal distributions with rate parameter ν. The goal of physicists is to construct confidence intervals for ν, and a discovery corresponds to a 5σ confidence interval that excludes $\nu = 0$. We model this problem by testing

$$H_0 : \nu = 0 \quad \text{versus} \quad H_1 : \nu \geq \pi \tag{mLFHT}$$

for fixed $\pi > 0$. In particular, a discovery can be claimed if H_0 is rejected. More precisely, given $C, \epsilon, R \geq 0$, let $\mathcal{P}_\mu(C, \epsilon, R)$ denote the set of triples (P_X, P_Y, P_Z) such that the following three conditions hold:

(*i*) P_X, P_Y and P_Z have μ-densities bounded by C,
(*ii*) $\mathrm{MMD}(P_X, P_Y) \geq \epsilon$ holds, and
(*iii*) $\mathrm{MMD}(P_Z, (1-\nu)P_X + \nu P_Y) \leq R \cdot \mathrm{MMD}(P_X, P_Y)$,

where we define $\nu = \nu(P_X, P_Y, P_Z) = \arg\min_{\nu' \in \mathbb{R}} \mathrm{MMD}(P_Z, (1-\nu')P_X + \nu' P_Y)$. For some $\pi > 0$, consider the two hypotheses

$$\begin{aligned} H_0 = \;& H_0(C, \epsilon, \pi, R) : (P_X, P_Y, P_Z) \in \mathcal{P}_\mu(C, \epsilon, R) \text{ and } \nu(P_X, P_Y, P_Z) = 0 \\ H_1 = \;& H_1(C, \epsilon, \pi, R) : (P_X, P_Y, P_Z) \in \mathcal{P}_\mu(C, \epsilon, R) \text{ and } \nu(P_X, P_Y, P_Z) \geq \pi. \end{aligned} \tag{1.6.1}$$

Notice that R controls the level of mis-specification in directions orthogonal to the line connecting the kernel embeddings of P_X and P_Y. Setting $R = 0$ simply asserts that P_Z is guaranteed to be a mixture of P_X and P_Y under both hypotheses and further taking $\pi = 1$ recovers LFHT. Given the parameters C, ϵ, π, R, mixed LFHT (mLFHT) is the problem of testing H_0 against H_1, as defined in (1.6.1), based on n, n, m observations from P_X, P_Y, P_Z respectively

1.6.2 A Kernel-Based Test for mLFHT

The following is a key result in understanding the behaviour of RKHSs and MMD, it is a consequence of the general fact that every compact self-adjoint operator is diagonalizable.

Theorem 1.6.1 (Hilbert–Schmidt theorem [173])**.** *Suppose that $K \in L^2(\mu \otimes \mu)$ is symmetric. Then there exists a sequence $(\lambda_j)_{j \geq 1} \in \ell^2$ and an orthonormal basis $\{e_j\}_{j \geq 1}$ of $L^2(\mu)$ such that $K(x, y) = \sum_{j \geq 1} \lambda_j e_j(x) e_j(y)$ for all $j \geq 1$, where convergence is in $L^2(\mu \otimes \mu)$.*

Throughout the rest of Section 1.6, statements involving $\mathcal{X}, \mu$ and K should implicitly be understood as holding for any choice of the three objects for which Theorem 1.6.1 holds. We warn the reader about the fact that given fixed $\mathcal{X}$ and K, the eigenvalues $\lambda_j \geq 0$ depend on the base measure μ that is chosen.

In our proofs we work with the kernel embedding of empirical measures for which we need to modify the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$, and thus MMD, by removing the diagonal terms. This debiases and reduces the variance of these empirical estimates. More concretely, given i.i.d.

samples X, Y of size n, m respectively and corresponding empirical measures $\widehat{P}_X, \widehat{P}_Y$, we define

$$\begin{aligned}\mathrm{MMD}_u^2(\widehat{P}_X, \widehat{P}_Y) \coloneqq \frac{1}{n(n-1)} \sum_{i \neq j} K(X_i, X_j) + \frac{1}{m(m-1)} \sum_{i \neq j} K(Y_i, Y_j) \\ - \frac{2}{mn} \sum_{i,j} K(X_i, Y_j).\end{aligned} \tag{1.6.2}$$

We also write $\langle \theta_{\widehat{P}_X}, \theta_{\widehat{P}_X} \rangle_{u,\mathcal{H}_K} \coloneqq \|\theta_{\widehat{P}_X}\|^2_{u,\mathcal{H}_K} \coloneqq \frac{1}{n(n-1)} \sum_{i \neq j} K(X_i, X_j)$ and extend linearly. The u stands for unbiased, since $\mathbb{E}\,\mathrm{MMD}_u^2(\widehat{P}_X, \widehat{P}_Y) = \mathrm{MMD}^2(P_X, P_Y) \neq \mathbb{E}\,\mathrm{MMD}^2(\widehat{P}_X, \widehat{P}_Y)$.

Suppose, as usual, that we have samples X, Y, Z of sizes n, n, m from the probability measures P_X, P_Y, P_Z and that our goal is to test between H_0 and H_1, defined in (1.6.1), for given parameters C, ϵ, π, R. Write $\widehat{P}_X$ for the empirical measure of sample X, and analogously for Y, Z. The core of our test statistic for (mLFHT) is the following:

$$T(X, Y, Z) \coloneqq \langle \theta_{\widehat{P}_Z}, \theta_{\widehat{P}_Y} - \theta_{\widehat{P}_X} \rangle_{u,\mathcal{H}_K} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \Big\{ K(Z_j, Y_i) - K(Z_j, X_i) \Big\}. \tag{1.6.3}$$

Further define

$$\gamma(X, Y, \pi) = \frac{\pi}{2} \mathrm{MMD}_u^2(\widehat{P}_X, \widehat{P}_Y) + T(X, Y, X). \tag{1.6.4}$$

The final output of our procedure is

$$\text{reject the null if } T(X, Y, Z) \geq \gamma(X, Y, \pi) \text{ and accept otherwise.} \tag{1.6.5}$$

The above has a natural geometric interpretation: we reject the null hypothesis if the projection of $\theta_{\widehat{P}_Z} - \theta_{\widehat{P}_X}$ onto the vector $\theta_{\widehat{P}_Y} - \theta_{\widehat{P}_X}$ falls further than $\pi/2$ along the segment joining $\theta_{\widehat{P}_X}$ to $\theta_{\widehat{P}_Y}$, up to deviations due to the deleted diagonal terms.

We are in a position to state our result on the minimax sample complexity of mLFHT under MMD separation. For simplicity we set the mis-specification parameter R to zero, for a more general statement of the result see Chapter 3.

Theorem 1.6.2. *Let us observe data* $X_1, \ldots, X_n \overset{iid}{\sim} P_X, Y_1, \ldots, Y_n \overset{iid}{\sim} P_Y$ *and* $Z_1, \ldots, Z_m \overset{iid}{\sim} P_Z$. *Then, there exists a finite universal constant* c *such that the procedure defined in* (1.6.5) *has total error bounded by* 5% *for testing* H_0 *vs* H_1, *as defined in* (1.6.1) *with* R *set to* 0, *provided*

$$\min\{m, n\} \geq c \frac{C\|\lambda\|_\infty}{\pi^2 \epsilon^2} \qquad \text{and} \qquad \min\{n, \sqrt{nm}\} \geq c \frac{C\|\lambda\|_2}{\pi \epsilon^2}.$$

The proof of Theorem 1.6.2 is a generalization of our analysis of the L^2-tester from Section 1.4.4, the strategy is simply to use the eigendecomposition available thanks to the Hilbert-Schmidt theorem. In addition to the new mixing parameter π, one difference from our previous results is that we put less restrictions on the class $\mathcal{P}_\mu$ where our distributions live. Indeed, our only requirement is that they have bounded density with respect to the base measure μ we choose, in contrast with the smoothness assumptions of the previous subsection.

The bounds on the sample sizes n and m depend on the sequence of eigenvalues λ, which in turn depends on the base measure μ that is chosen. Therefore, given a fixed kernel K, it would be possible to try to optimize μ so that the resulting bounds in Theorem 1.6.2 are as favorable as possible. However, since μ is also present in the definition of the hypothesis class $\mathcal{P}_\mu$ itself, it seems difficult to disentangle the relationships.

We also obtain a partial converse to Theorem 1.6.2 in the form of a minimax lower bound. Due to the overly general setting of the problem, it harder to construct the least favorable distributions. Therefore, we need additional assumptions on the problem.

Theorem 1.6.3. *Given $J \geq 2$, let $\|\lambda\|_{2J}^2 := \sum_{j=2}^J \lambda_j^2$. Let*

$$J_\epsilon^\star := \max\Big\{J : \sup_{\eta_j = \pm 1} \Big\| \sum_{j=2}^J \eta_j \sqrt{\lambda_j} e_j \Big\|_\infty \leq \frac{\|\lambda\|_{2J}}{2\epsilon}\Big\}.$$

Consider the setting of Theorem 1.6.2 and suppose that $\int_\mathcal{X} K(x,y)\mu(\mathrm{d}x) \equiv \lambda_1$, $\mu(\mathcal{X}) = 1$ and $\sup_{x\in\mathcal{X}} K(x,x) \leq 1$. There exists a universal constant $c > 0$ such that any test with total error at most 5% must use

$$m \geq \frac{c\lambda_2}{\pi^2\epsilon^2} \quad \text{and} \quad n \geq c\frac{\|\lambda\|_{2J_\epsilon^\star}}{\epsilon^2} \quad \text{and} \quad \max\{\pi m, \sqrt{mn}\} \geq c\frac{\|\lambda\|_{2J_\epsilon^\star}}{\pi\epsilon^2}.$$

The requirements $\sup_{x\in\mathcal{X}} K(x,x) \leq 1$ and $\mu(\mathcal{X}) = 1$ are is essentially without loss of generality, as μ and K can be rescaled. The condition $\int_\mathcal{X} K(x,y)\mu(\mathrm{d}x) \equiv \lambda_1$ implies that the top eigenfunction e_1 is equal to a constant or equivalently, that $K(\mathrm{d}x, y)\mu(\mathrm{d}x)$ defines a Markov kernel for μ-almost every $y \in \mathcal{X}$, up to a normalizing constant.

The lower bound construction is inspired by that for $\mathcal{P}_{\mathsf{Db}}$ and $\mathcal{P}_{\mathsf{H}}$, which we discuss in more detail in Section 1.7.1. Simply take a vector of independent uniformly random signs $\eta = (\eta_1, \eta_2 \dots) \in \{\pm 1\}^\mathbb{N}$ and define the μ-density $f_\eta \propto 1 + \rho\sum_{j=2}^J \eta_j e_j$ for $J \geq 2$ and some scaling factor $\rho > 0$ chosen so that $\mathrm{MMD}(\mathrm{unif}(\mathcal{X}), f_\eta) \geq \epsilon$ for any realization of the signs η. Recall that e_j are the orthogonal eigenfunctions of the kernel K, so that $\int e_j(x)\mu(\mathrm{d}x) = \langle e_1, e_j\rangle_{L^2(\mu)} = 0$, which implies that f_η is a valid probability density with respect to μ provided it is non-negative. This condition is precisely the reason why we must take a potentially finite cutoff J: we are unable to control the size of the perturbation $\|\sum_{j=1}^J \eta_j e_j\|_\infty$ for $J = \infty$. The largest J take we are able to take while still guaranteeing non-negativity of the construction f_η is precisely $J = J_\epsilon^\star$. Once we have the construction in hand, the derivation of the lower bound follows in a completely analogous fashion to Section 1.7 and Paninski's construction in particular.

An apparent weakness of Theorem 1.6.3 is its reliance on the unknown value $J_\epsilon^\star$, which depends on the specifics of the kernel K and base measure μ. Determining it is potentially nontrivial even for simple kernels. Slightly weakening Theorem 1.6.3 we obtain the following corollary, which shows that the dependence on $\|\lambda\|_2$ is tight, at least for small enough ϵ.

Corollary 1.6.4. *Suppose $J \geq 2$ is such that $\sum_{j=2}^J \lambda_j^2 \geq c^2\|\lambda\|_2^2$ for some $c \leq 1$. Then $\|\lambda\|_{2J_\epsilon^\star}$ can be replaced by $c\|\lambda\|_2$ in Theorem 1.6.3 whenever $\epsilon \leq c\|\lambda\|_2/(2\sqrt{J-1})$.*

The correct dependence on the signal rate π is the most pressing question left open by our theoretical results.

1.6.3 Learning the Kernel and the Empirical Trade-Off

Given a fixed kernel K, our Theorems 1.6.2 and 1.6.3 show that the sample complexity depends on the separation ϵ under the given MMD as well as the spectrum $\lambda = \lambda(\mu, K)$ of the kernel. Thus, to have good test performance we need to use a kernel K that is well-adapted to the problem at hand. In practice, however, instead of using a fixed kernel it would be only natural to use part of the simulated data to try to learn a good kernel. Due to the resulting dependence between the data and the kernel, Theorems 1.6.2 and 1.6.3 don't apply anymore. Our main experimental contribution is to confirm the existence of an asymmetric simulation-experimentation trade-off similar to Figure 1.1, that appears in spite of said dependence.

Training Objective

Consider taking a part of the simulation data and setting it aside for training the kernel; call this $(X^{\mathsf{tr}}, Y^{\mathsf{tr}})$. Writing $\widehat{P}_{X^{\mathsf{tr}}}, \widehat{P}_{Y^{\mathsf{tr}}}$ for their empirical measures, we maximize the objective

$$\widehat{J}(X^{\mathsf{tr}}, Y^{\mathsf{tr}}; K) = \frac{\mathrm{MMD}_u^2(\widehat{P}_{X^{\mathsf{tr}}}, \widehat{P}_{Y^{\mathsf{tr}}}; K)}{\widehat{\sigma}(X^{\mathsf{tr}}, Y^{\mathsf{tr}}; K)}, \tag{1.6.6}$$

which was introduced in [187], originally for two-sample testing. Here $\widehat{\sigma}^2$ is an estimator of the variance of $\mathrm{MMD}_u^2(\widehat{P}_{X^{\mathsf{tr}}}, \widehat{P}_{Y^{\mathsf{tr}}}; K)$ defined in [143] as follows. For each pair i, j define

$$H_{ij} := K(X_i^{\mathsf{tr}}, X_j^{\mathsf{tr}}) + K(Y_i^{\mathsf{tr}}, Y_j^{\mathsf{tr}}) - K(X_i^{\mathsf{tr}}, Y_j^{\mathsf{tr}}) - K(Y_i^{\mathsf{tr}}, X_j^{\mathsf{tr}}). \tag{1.6.7}$$

The variance estimate is then computed via

$$\widehat{\sigma}^2(X^{n_{\mathsf{tr}}}, Y^{n_{\mathsf{tr}}}; K) = \frac{4}{n_{\mathsf{tr}}^3} \sum_{i=1}^{n_{\mathsf{tr}}} \left(\sum_{j=1}^{n_{\mathsf{tr}}} H_{ij} \right)^2 - \frac{4}{n_{\mathsf{tr}}^4} \left(\sum_{i=1}^{n_{\mathsf{tr}}} \sum_{j=1}^{n_{\mathsf{tr}}} H_{ij} \right)^2. \tag{1.6.8}$$

Intuitively, the objective J aims to separate P_X from P_Y while keeping variance low. We optimize the objective (1.6.6) over the choice of kernel K, which we parametrize as follows. For $\tau \in (0, 1)$ we let

$$K(x, y) = \big((1-\tau) G_\sigma(\varphi_\omega(x), \varphi_\omega(y)) + \tau\big) \cdot G_{\sigma_0}(x + \varphi_{\omega'}(x), y + \varphi_{\omega'}(y)),$$

where G_σ is the Gaussian kernel with variance σ^2; $\varphi_\omega, \varphi_{\omega'}$ are neural networks with parameters ω, ω', and $\sigma, \sigma_0, \tau, \omega, \omega'$ are all trainable scalars. This is the same architecture introduced in [143] and is the best performing one in our experiments.

Experiment I : Image Source Detection

Our first empirical study looks at the task of detecting whether images come from the CIFAR-10 [136] dataset or a SOTA generative model (DDPM) [95, 167]. While source detection is on its own interesting, it turns out that detecting whether a group of images comes from the generative model versus the real dataset can be too "easy" (see experiments in [118]). Therefore, we consider a mixed alternative, where the alternative hypothesis is not simply

the generative model but CIFAR with planted DDPM images. Namely, our n labeled images come from the following distributions:

$$P_X = \text{CIFAR}, \quad \text{and} \quad P_Y = \pi \cdot \text{DDPM} + (1 - \pi) \cdot \text{CIFAR}. \tag{1.6.9}$$

The goal is to test whether the fraction of the m unlabeled observations Z which are fakes generated by DDPM exceeds π.

Experiment II : Higgs Boson Dataset

The 2012 discovery of the Higgs boson by the ATLAS and CMS experiments [1, 40] marked a significant milestone in physics. The statistical problem inherent in the experiment is well-modeled by (mLFHT), using a signal rate π predicted by theory and misspecification parameter $R = 0$. We use the open source Higgs dataset available at http://archive.ics.uci.edu/ml/datasets/HIGGS and record the probability of error over multiple kernel training runs with different sample sizes m, n.

The Empirical Trade-Off

Figure 1.4 shows the error probability of the two experiments described above, as we vary the sample sizes m and n. Note that here n also includes those simulation samples that are used for training the neural network, so that our minimax theory doesn't apply. However, we can clearly see from both plots that there is an asymmetric trade-off between the number of simulation samples n and real data m. In particular, for a given level of error, one needs fewer real data samples m as we take $n \to \infty$ compared to the opposite case.

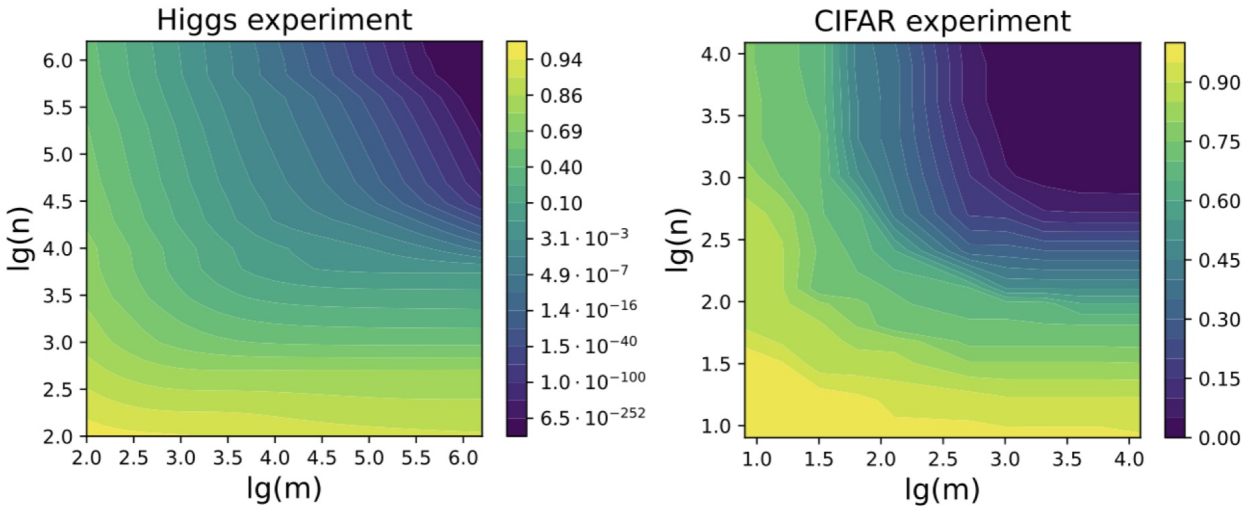


Figure 1.4: n versus m trade-off for the Higgs and CIFAR experiments using our test. Error probabilities are estimated by normal approximation for Higgs and simulated for CIFAR.

1.7 Minimax Lower Bounds for LFHT

In this section we outline our approach to proving matching minimax lower bounds for LFHT, and testing problems more generally. Suppose we observe a sample X from an unknown distribution Q, and our goal is to test between the two hypotheses

$$H_0 : Q \in \mathcal{Q}_0 \qquad \text{versus} \qquad H_1 : Q \in \mathcal{Q}_1,$$

where $\mathcal{Q}_i, i = 0, 1$ are disjoint sets of probability distributions. Tsybakov's [198] method of fuzzy hypotheses gives us a recipe for deriving a lower bound on the minimax error probability of this problem. Let Q_i be random probability distributions supported on $\mathcal{Q}_i$ respectively. Then

$$\text{best worst-case error} = \inf_{\Psi} \max_{i=0,1} \sup_{Q \in \mathcal{Q}_i} \mathbb{P}_{X \sim Q}(\Psi(X) \neq i) \geq \frac{1}{2}(1 - \mathsf{TV}(\mathbb{E}Q_0, \mathbb{E}Q_1)) \tag{1.7.1}$$

holds, which is proven by replacing the supremum over $\mathcal{Q}_i$ with an average. Therefore, the problem is reduced to bounding the total variation distance between two mixtures. There are many inequalities at our disposal, but the most useful for us is

$$1 - \mathsf{TV}(P, Q) \geq \frac{1/2}{1 + \chi^2(P\|Q)}, \tag{1.7.2}$$

which is valid for any P and Q and can be found for example in [168, Section 7.6].

So that we may give concrete examples, let us once again focus on the case of discrete distributions for our discussion. There are two famous constructions for discrete distributions that we cover here.

1.7.1 Perturbation of the Uniform Distribution

The first construction is attributed to Paninski [162] in the computer science literature, however the idea is quite old and appears in [110] for the exact same purpose. Given a desired separation $\epsilon \in (0, 1)$, an even support size k and a sequence of signs $\eta_1, \ldots, \eta_{k/2} \in \{\pm 1\}$ we define the pmf p_η by

$$p_\eta(2i) = \frac{1 + \eta_i \epsilon}{k} = \frac{2}{k} - p_\eta(2i - 1)$$

for all $i \in [k/2]$. Also define $p_0 = (1/k, \ldots, 1/k)$ to be the uniform distribution. Taking the signs η to be uniformly random, one can immediately derive minimax sample complexity lower bounds for goodness-of-fit testing, and consequently for two-sample testing and LFHT by reduction c.f. Section 1.4.2, by computing

$$\begin{aligned} 1 + \chi^2(\mathbb{E}_\eta p_\eta^{\otimes n} \| p_0^{\otimes n}) &= \sum_{x_1, \ldots, x_n \in [k]^n} \frac{\left(\mathbb{E}_\eta \prod_{i=1}^n p_\eta(x_i)\right)^2}{\prod_{i=1}^n p_0(x_i)} \\ &= k^n \mathbb{E}_{\eta, \eta'} \left[\left(\sum_{x=1}^k p_\eta(x) p_{\eta'}(x) \right)^n \right], \end{aligned}$$

where η, η' are independent and identically distributed. Notice now that

$$\sum_{x=1}^{k} p_\eta(x) p_{\eta'}(x) = \frac{1}{k}\left(1 + \frac{2\epsilon^2}{k}\langle \eta, \eta' \rangle\right).$$

Plugging into our bound of the χ^2-divergence, and using the inequalities $1 + x \leq \exp(x)$ and $\cosh(x) \leq \exp(x^2/2)$ valid for all $x \in \mathbb{R}$, we obtain

$$\begin{aligned} 1 + \chi^2(\mathbb{E}_\eta p_\eta^{\otimes n} \| p_0^{\otimes n}) &\leq \mathbb{E}_{\eta,\eta'} \exp\left(\frac{2\epsilon^2 n \langle \eta, \eta' \rangle}{k}\right) \\ &= \left(\mathbb{E}_{\eta_1,\eta_1'} \exp\left(\frac{2\epsilon^2 n \eta_1 \eta_1'}{k}\right)\right)^{k/2} \\ &= \cosh^{k/2}\left(\frac{2\epsilon^2 n}{k}\right) \\ &\leq \exp\left(\frac{\epsilon^4 n^2}{k}\right). \end{aligned} \tag{1.7.3}$$

Via (1.7.1) and (1.7.2) this immediately implies the familiar lower bound $\sqrt{k \log(1/\delta)}/\epsilon^2$ on the sample complexity of goodness-of-fit testing. The idea to bound the total variation distance between mixtures of products using the χ^2-divergence, as above, is attributed to Ingster (see for example [110, 113]).

The construction can also be used to derive optimal lower bounds for our problem LFHT. For this we need to utilize the chain rule of the χ^2-divergence. By conditioning on the outcome of the X-sample, we have

$$\chi^2(\mathbb{E} p_\eta^{\otimes n} \otimes p_0^{\otimes n} \otimes p_\eta^{\otimes m} \| \mathbb{E} p_\eta^{\otimes n} \otimes p_0^{\otimes (n+m)}) = \mathbb{E}_X \chi^2((Z_1, \ldots, Z_m) \| p_0^{\otimes m} | X),$$

which uses the fact that $X = (X_1, \ldots, X_n)$ has the same marginal distribution under both hypotheses. If we expand the square in the conditional χ^2-divergence we obtain an expression similar to (1.7.3), but now η, η' are drawn independently from the posterior given X, and are no longer unconditionally independent. The remainder of the computation is similar to the one above, the final bound we eventually obtain is $\exp(c(nm + n^2)\epsilon^4/k) - 1$ for some constant c, which leads to the sample complexity lower bound $mn \gtrsim k \log(1/\delta)/\epsilon^4$.

1.7.2 Valiant's Construction

The construction in Section 1.7.1 along with the reductions in Section 1.4.2 are sufficient to give optimal sample complexity lower bounds for LFHT, except for the class of all discrete distributions $\mathcal{P}_{\mathsf{D}}$. In the regime $1/\epsilon^4 \leq k$ this class requires a new construction. This is the same threshold where the sample complexity of two-sample testing, given by $\frac{\sqrt{k}}{\epsilon^4} + \frac{k^{2/3}}{\epsilon^{4/3}}$ in the constant error regime, undergoes a phase transition. In fact, the same construction that shows the matching lower bounds for two-sample testing can be adapted to the LFHT setting, which we detail below.

Instead of the i.i.d. sampling model we use the Poissonized model and rely on the formalism of pseudo-distributions as described in [62]. Specifically, suppose we can construct a random

vector $(p, q) \in [0,1]^2$ such that 1) $\mathbb{E}p = \mathbb{E}q = \Theta(1/k)$ and $\mathbb{E}|p-q| = \Theta(\epsilon/k)$; and 2) one of the following χ^2 upper bounds hold:

$$\chi^2\Big(\mathbb{E}\big[\operatorname{Poi}(np)\otimes\operatorname{Poi}(nq)\otimes\operatorname{Poi}(mp)\big]\Big\|\mathbb{E}\big[\operatorname{Poi}(np)\otimes\operatorname{Poi}(nq)\otimes\operatorname{Poi}(mq)\big]\Big) \le B(n,m,\epsilon,k)$$

or

$$\chi^2\Big(\mathbb{E}\big[\operatorname{Poi}(nq)\otimes\operatorname{Poi}(np)\otimes\operatorname{Poi}(mp)\big]\Big\|\mathbb{E}\big[\operatorname{Poi}(np)\otimes\operatorname{Poi}(nq)\otimes\operatorname{Poi}(mp)\big]\Big) \le B(n,m,\epsilon,k);$$

then $(n,m) \in \mathcal{R}_{\mathsf{LF}}(\epsilon,\delta,\mathcal{P}_{\mathsf{D}})$ requires $kB(n,m,\epsilon,k) \gtrsim \log(1/\delta)$; this is essentially via (1.7.2).

Remark 11. *The classical approach for proving lower bounds in the 'sparse' regime (i.e. when the number of observations n is less than the support size k) is due to Paul Valiant. His approach was developed for testing symmetric properties and is conceptually considerably more involved than the direct calculation we outline below. Valiant's method didn't use pseudo-distributions and instead relied on the notion of fingerprints, which is equivalent to randomly permuting the support, and Roos' inequality [177] to compare fingerprints with a product of Poisson distributions. He calls the main inequality, which bounds the total variation of the construction via difference in moments, the "wishful thinking theorem" [200, Theorem 6]. This has been used as a black box for proving lower bounds with great success, for example for two-sample testing [39]. However, newer work shows how one can prove the same lower bounds via the same constructions using a more direct approach by directly bounding the mutual information [64] and most recently the KL-divergence [62]. The latter work in fact passed from the KL-divergence to the more amenable χ^2-divergence without realizing [62, Lemma 5.15]. This prompts us to simply bound the χ^2-divergence directly and thus reduce the entire lower bound program to an elementary computation.*

The $m \le n \le k$ Case

Let p, q be two random variables defined as

$$(p,q) = \begin{cases} (\frac{1}{n}, \frac{1}{n}) & \text{with probability } \frac{n}{k}, \\ (\frac{\epsilon}{k}, \frac{2\epsilon}{k}) & \text{with probability } \frac{1}{2}(1-\frac{n}{k}), \\ (\frac{\epsilon}{k}, 0) & \text{with probability } \frac{1}{2}(1-\frac{n}{k}). \end{cases}$$

Note that $\mathbb{E}[p] = \mathbb{E}[q] = \Theta(1/k)$ and $\mathbb{E}|p-q| = \Theta(\epsilon/k)$, so they form valid ϵ-TV separated pseudo-distributions. Define the random variables $U, V \in \mathbb{R}^3$ with distribution given by

$$U|(p,q) \sim \operatorname{Poi}(np)\otimes\operatorname{Poi}(nq)\otimes\operatorname{Poi}(mp),$$
$$V|(p,q) \sim \operatorname{Poi}(np)\otimes\operatorname{Poi}(nq)\otimes\operatorname{Poi}(mq).$$

It is not hard to verify that for any $(a,b,c) \in \mathbb{N}^3$ the following two estimates hold

$$\mathbb{P}(V=(a,b,c)) = \Omega\left(\frac{1}{a!b!c!}\right)\begin{cases} 1 & \text{if } (a,b,c)=(0,0,0), \\ \frac{n}{k}\left(\frac{m}{n}\right)^c & \text{otherwise, and} \end{cases}$$

$$\Big|\mathbb{P}(U=(a,b,c)) - \mathbb{P}(V=(a,b,c))\Big| = \frac{\Theta(1)}{a!b!c!}\left(\frac{\epsilon}{k}\right)^{a+b+c} n^{a+b} m^c \begin{cases} \frac{nm\epsilon^2}{k^2} & \text{if } b=c=0, \\ \frac{2^b m\epsilon}{k} & \text{if } b\ge 1, c=0, \\ \frac{n\epsilon}{k} & \text{if } b=0, c=1, \\ 2^{b+c} & \text{otherwise.} \end{cases}$$

With these we are ready to bound the χ^2-divergence between the distributions of U and V. Simply plugging in the estimates and checking the special cases that arise, one can show that

$$\chi^2(U\|V) = \sum_{(a,b,c)\in\mathbb{N}^3} \frac{(\mathbb{P}(U=(a,b,c)) - \mathbb{P}(V=(a,b,c)))^2}{\mathbb{P}(V=(a,b,c))} \lesssim \frac{\epsilon^4 mn^2}{k^3}.$$

As explained in the introduction of this section, via [62] this implies the minimax sample complexity lower bound $mn^2 \gtrsim \log(1/\delta)k^2/\epsilon^4$.

The $n \le m \le k$ Case

The approach in this case is very similar to the $m \le n \le k$ case with one important distinction. Let p, q be two random variables defined as

$$(p,q) = \begin{cases} (\frac{1}{m}, \frac{1}{m}) & \text{with probability } \frac{m}{k}, \\ (\frac{\epsilon}{k}, \frac{2\epsilon}{k}) & \text{with probability } \frac{1}{2}(1-\frac{m}{k}), \\ (\frac{\epsilon}{k}, 0) & \text{with probability } \frac{1}{2}(1-\frac{m}{k}). \end{cases}$$

Let $U, V \in \mathbb{R}^3$ be random, whose distribution is given by

$$U|(p,q) \sim \mathrm{Poi}(np) \otimes \mathrm{Poi}(nq) \otimes \mathrm{Poi}(mp),$$
$$V|(p,q) \sim \mathrm{Poi}(nq) \otimes \mathrm{Poi}(np) \otimes \mathrm{Poi}(mp).$$

Estimates for $\mathbb{P}(V=(a,b,c))$ and $\mathbb{P}(V=(a,b,c)) - \mathbb{P}(V=(a,b,c))$ that are analogous to the $n \le m \le k$ case above yield the final bound $\chi^2(U\|V) \lesssim \frac{\epsilon^4 mn^2}{k^3}$, which in turn implies the desired bound $n^2 m \gtrsim \log(1/\delta)k^2/\epsilon^4$.

Notice the difference in the definition of $(U,V)|(p,q)$ between the cases $m \le n \le k$ and $n \le m \le k$. In the former case the third marginal of U and V are different, while in the latter case the third marginal of U and V are the same. This is because m is larger than n and if we let U and V deviate on the third marginal, which has "sample size" m, we would reveal too much about U and V and make distinguishing the two distributions too easy. If one were to work trough the computation it would result in the bound $\chi^2(U\|V) \lesssim \epsilon^4 nm^2/k^3$, which is sub-optimal. The reverse argument applies to the case $m \le n \le k$: we want to put the difference of U, V on the third marginal as that provides less information compared to the first two.

Chapter 2

Likelihood-Free Hypothesis Testing

This chapter is a reproduction of [77], which is joint work with Yury Polyanskiy.

2.1 Introduction

A setting that we call *likelihood-free inference (LFI)*, also known as simulation based inference (SBI), has independently emerged in many areas of science over the past decades. Given an expensive to collect "experimental" dataset and the ability to simulate from a high fidelity, often mechanistic, stochastic model, whose output distribution and likelihood is intractable and inapproximable, how does one perform model selection, parameter estimation or construct confidence sets? The list of disciplines where such highly complex black-box simulators are used is long, and include particle physics, astrophysics, climate science, epidemiology, neuroscience and ecology to just name a few. For some of the above fields, such as climate modeling, the bottleneck resource is in fact the simulated data as opposed to the experimental data. In either case, understanding the trade-off between the number of simulations and experiments necessary to do valid inference is crucial. Our aim in this paper is to introduce a theoretical framework under which LFI can be studied using the tools of non-parametric statistics and information theory.

To illustrate we draw an example from high energy physics, where LFI methods are used and developed extensively. The discovery of the Higgs boson in 2012 [40, 5] is regarded as the crowning achievement of the Large Hadron collider (LHC) - the most expensive instrument ever built. Using a composition of complex simulators [6, 73, 48, 183, 8] modeling the standard model and the detection process, physicists are able to simulate the results of LHC experiments. Given actual data $Z_1, \dots, Z_m$ from the collider, to verify existence of the Higgs boson one tests whether the null hypothesis (physics without the Higgs boson, or $Z_i \overset{iid}{\sim} \mathbb{P}_0$) or the alternative hypothesis (physics with the Higgs boson, or $Z_i \overset{iid}{\sim} \mathbb{P}_1$) describes the experimental data more accurately. The standard Neyman-Pearson likelihood ratio test is not implementable since $\mathbb{P}_0$ and $\mathbb{P}_1$ are only available via simulators.

How was this statistical test actually performed? First, a probabilistic classifier C was trained on simulated data to distinguish the two hypotheses (a boosted decision tree to be more specific). Then, the proportion of real data points falling in the set $S = \{x \in \mathbb{R}^d : C(x) \leq t\}$ was computed, where t is chosen to maximize an asymptotic approximation of the power.

Finally, p-values are reported based on the asymptotic distribution under a Poisson sampling model [49, 142]. Summarizing, the "Higgs boson" test was performing the simple comparison

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{Z_i \in S\} \lessgtr \gamma, \tag{Scheffé}$$

where $Z_1, \ldots, Z_m$ are the real data and γ is some threshold. Such count-based tests, named after Scheffé in folklore [60, Section 6], are quite intuitive.

Notice that Scheffé's test converts each observation Z_i into a binary 0/1 value. This extreme quantization certainly helps robustness, but should raise the suspicion of potential loss of power. Indeed, when the distributions under both hypotheses are completely known, the optimal Neyman-Pearson test thresholds the sum of *real-valued* logarithms of the likelihood-ratio. Thus, it is natural to expect that a good test should aggregate non-binary values. This is what motivated this work originally, although follow-up work [74] has shown that Scheffé's test with a properly trained classifier can also be optimal.

Let us describe the test that we study for most of this paper. Given estimates $\widehat{p}_0, \widehat{p}_1$ of the density of the null and alternative distributions based on simulated samples, our test proceeds via the comparison

$$\frac{2}{m}\sum_{i=1}^{m}(\widehat{p}_0(Z_i) - \widehat{p}_1(Z_i)) \lessgtr \gamma \tag{2.1.1}$$

where $Z_1, \ldots, Z_m$ are the real data. Tests of this kind originate from the famous goodness-of-fit work of Ingster [110], which corresponds to taking $\widehat{p}_0 = p_0$, as the null-density is known exactly.[1] The surprising observation of Ingster was that such a test is able to reject the null hypothesis that $Z_i \overset{iid}{\sim} p_0$ even when the true distribution of Z is much closer to p_0 than described by the optimal density-estimation rate; in other words *goodness-of-fit testing is significantly easier than estimation*. In fact we will use $\gamma = \|\widehat{p}_0\|_2^2 - \|\widehat{p}_1\|_2^2$ in which case (2.1.1) boils down to the comparison of two squared L^2-distances.

Our overall goal is to understand the trade-off between the number n of simulated observations and the size of the actual data set m. The characterization of this tradeoff is reminiscent of the rate-regions in multi-user information theory, but there is an important difference that we wanted to emphasize for the reader. In information theory, the problem is most often stated in the form "given a distribution $P_{X,Y,Z}$, or a channel $P_{Y,Z|X}$, find the rate region", with the distribution being completely specified ahead of time. In minimax statistics, however, distributions are a priori only known to belong to a certain class. In *estimation problems* the fundamental limits are thus defined by minimizing the estimation error over this class, and the theoretical goal is to characterize the worst-case rate at which this error converges to zero as the sample size grows to infinity. The definition of the fundamental limit in *testing problems*, however, is more subtle. If the total variation separation ϵ between the null and alternative distribution is fixed, and the number of samples is taken to infinity, then the rate of convergence trivializes and becomes exponentially decreasing in n. By now a

[1] In the case of discrete distributions on a finite (but large) alphabet, the idea was rediscovered by the computer science community startin with [81]. Moreover, the difference of L^2-norms statistic was first studied in [123]. See Section 2.1.2 for more on the latter.

standard definition of fundamental limit, as suggested by Ingster following ideas of Pittman efficiency, is to vary ϵ with n and to find the fastest possible decrease of ϵ so as to still have an acceptable probability of error. This is the approach taken in the literature on goodness-of-fit and two-sample testing, and also the one we adopt here. This perspective is also widely used in TCS where the optimal value of n, as a function of ϵ, is referred to as the "sample complexity" of the problem.

Specifically, we assume that it is known a priori that the two distributions $\mathbb{P}_0, \mathbb{P}_1$ belong to a known class $\mathcal{P}$ and are ϵ-separated under total variation. Given a large number n of samples simulated from $\mathbb{P}_0$ and $\mathbb{P}_1$ and m samples $Z_1, \ldots, Z_m$ from the experiment, our goal is to test which of the $\mathbb{P}_i$ generated the data. If n is sufficiently large to estimate $\mathbb{P}_i$ in total variation to precision $\epsilon/10$, then one can perform the hypothesis test with $m \asymp 1/\epsilon^2$ experimental samples, which is information-theoretically optimal even under oracle knowledge of $\mathbb{P}_i$'s. However, looking at the test (Scheffé) one may wonder if the full estimation of the distributions $\mathbb{P}_i$ is needed, or whether perhaps a suitable decision boundary could be found with a lot fewer simulated samples n. Unfortunately, our *first main result* disproves this intuition: *any test using the minimal $m \asymp 1/\epsilon^2$ dataset size will require n so large as to be enough to estimate the distributions of $\mathbb{P}_0$ and $\mathbb{P}_1$ to within accuracy $\asymp \epsilon$*, which is the distance separating the two hypotheses. In particular, any method minimizing m performs no different in the worst case, than pairing off-the-shelf density estimators $\widehat{p}_0, \widehat{p}_1$ and applying (Scheffé) with $S = \{\widehat{p}_1 \geq \widehat{p}_0\}$.

This result appears rather pessimistic and seems to invalidate the whole attraction of LFI, which after all hopes to circumvent the exorbitant number of simulation samples required for fully learning high-dimensional distributions. Fortunately, *our second result* offers a resolution: if more data samples $m \gg 1/\epsilon^2$ are collected, then testing is possible with n much smaller than required for density estimation. More precisely, when neither p_0 nor p_1 are known except through n i.i.d. samples from each, the test (2.1.1) is able to detect which of the two distributions generated the Z-sample, *even when the number of samples n is insufficient for any estimate $\widehat{p}_i$ to be within a distance $\asymp \epsilon = \mathsf{TV}(p_0, p_1)$* from the true values. In other words, the test is able to reliably detect the true hypotheses even though the estimates $\widehat{p}_i$ themselves have accuracy that is orders of magnitude larger than the separation ϵ between the hypotheses.

In summary, this paper shows that likelihood-free hypothesis testing (LFHT) is possible without learning the densities when $m \gg 1/\epsilon^2$, but not otherwise. It turns out that (appropriate analogues of) the simple test (2.1.1) has minimax optimal sample complexity up to constants in both n and m in all "regular" settings, see also the discussion at the end of Section 2.2.2.

2.1.1 Informal statement of the main result

Let us formulate the problem using the notation used throughout the rest of the paper. Suppose that we observe true data $Z \sim \mathbb{P}_{\mathsf{Z}}^{\otimes m}$ and that we have two candidate parameter settings for our simulator, from which we generate two artificial datasets $X \sim \mathbb{P}_{\mathsf{X}}^{\otimes n}$ and $Y \sim \mathbb{P}_{\mathsf{Y}}^{\otimes n}$. If we are convinced that one of the settings accurately reflects reality, we are faced with the problem of testing the hypothesis

$$H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_{\mathsf{Z}} \qquad \text{versus} \qquad H_1 : \mathbb{P}_{\mathsf{Y}} = \mathbb{P}_{\mathsf{Z}}. \tag{2.1.2}$$

Remark 12. *We emphasize that $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ are known only through the n simulated samples. Thus, (2.1.2) can be interpreted as binary hypothesis testing with approximately specified hypotheses. Alternatively, using the language of machine learning, we may think of this problem as having n labeled samples from both classes, and m unlabeled samples. The twist is that the unlabeled samples are guaranteed to have the same common label, that is, they all come from a single class. One can think of many examples of this setting occurring in genetic, medical and other studies.*

To put (2.1.2) in a minimax framework, suppose that $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P}$ for a known class $\mathcal{P}$, and that $\mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \geq \epsilon$. Clearly (2.1.2) becomes "easier" if we have a lot of data (large sample sizes n and m) or if the hypotheses are well-separated (large ϵ). We are interested in characterizing the pairs of values (n, m) as functions of ϵ and $\mathcal{P}$, for which the hypothesis test (2.1.2) can be performed with constant type-I and type-II error. Letting $n_{\mathsf{GoF}}(\epsilon, \mathcal{P})$ denote the minimax sample complexity of goodness-of-fit testing (Definition 2), we show for *several different classes* of $\mathcal{P}$, that (2.1.2) is possible with total error, say, 5% if and only if

$$m \gtrsim 1/\epsilon^2 \qquad \text{and} \qquad n \gtrsim n_{\mathsf{GoF}} \qquad \text{and} \qquad mn \gtrsim n_{\mathsf{GoF}}^2.$$

We also make the observation that $n_{\mathsf{GoF}}^2 \, \epsilon^2 \asymp n_{\mathsf{Est}}$ for these classes, where $n_{\mathsf{Est}}(\epsilon, \mathcal{P})$ denotes the minimax complexity of density estimation to ϵ-accuracy (Definition 4) with respect to total variation. This provides additional meaning to the mysterious formula of Ingster [110] for the sample complexity of goodness-of-fit testing over the class of β-smooth densities over $[0,1]^d$, see Table 2.1 below.[2] More importantly, it allows us to interpret (2.1.2) as an "interpolation" between different fundamental statistical procedures, namely

- $\mathsf{A} \leftrightarrow$ Binary hypothesis testing,
- $\mathsf{B} \leftrightarrow$ Estimation followed by robust binary hypothesis testing,
- $\mathsf{C} \leftrightarrow$ Two-sample testing,
- $\mathsf{D} \leftrightarrow$ Goodness-of-fit testing,

corresponding to the extreme points $\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}$ on Figure 2.1.

2.1.2 Related work

LHFT as defined in (2.1.2) initially appeared in Gutman's paper [90], building on Ziv's work [212], where the problem is studied for distributions on a fixed, finite alphabet. Ziv called the problem *classification with empirically observed statistics*, to emphasize the fact that hypotheses are specified only in terms of samples and the underlying true distributions are unknown. In [209] it is shown that the error exponent of Gutman's test is second order optimal. Recent work [98, 93, 92, 30] extends this problem to distributed and sequential

[2] A possible reason for this observation having been missed previously is that fundamental limits in statistics are usually presented in the form of *rates* of loss decrease with n, for example $r_{\mathsf{Est}}(n) =: n_{\mathsf{Est}}^{-1}(n) = 1/n^{\beta/(2\beta+d)}$ and $r_{\mathsf{GoF}}(n) =: n_{\mathsf{GoF}}^{-1}(n) = 1/n^{\beta/(2\beta+d/2)}$ for β-smooth densities. Unlike $n_{\mathsf{Est}} \asymp n_{\mathsf{GoF}}^2 \epsilon^2$ there seems to be no simple relation between r_{Est} and r_{GoF}.

testing. However, the setting of these papers is fundamentally different from ours, a point which we expand on below.

Given two arbitrary, unknown $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}$ over a finite alphabet of fixed size, Gutman's test (see [209, Equation (4)]) rejects the null hypothesis $H_0 : \mathbb{P}_{\mathsf{Z}} = \mathbb{P}_{\mathsf{X}}$ in favor of the alternative $H_1 : \mathbb{P}_{\mathsf{Z}} = \mathbb{P}_{\mathsf{Y}}$ if the statistic $\mathrm{GJS}(\widehat{\mathbb{P}}_{\mathsf{X}}, \widehat{\mathbb{P}}_{\mathsf{Z}}, \alpha)$ is large, where $\widehat{\mathbb{P}}$ denotes empirical measures, GJS denotes the generalized Jensen-Shannon divergence defined in [209, Equation (3)] and $\alpha = n/m$. In other words, it simply performs a two-sample test using the samples from $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Z}}$ of size n and m respectively, and completely discards the sample from $\mathbb{P}_{\mathsf{Y}}$. In light of our sample complexity results this is strictly sub-optimal due to minimax lower bounds on two-sample testing, see the difference of light gray and striped regions in Figure 2.1.

More generally, the method of types, which is a crucial tool for the works cited above, cannot be used to derive our results, because in the regime where the alphabet size k scales with the sample size n, the usual $\binom{n}{k} = e^{o(n)}$ approximation no longer holds, i.e. these factors affect estimation rates and do not lead to tight minimax results. As a consequence, one cannot deduce results about the minimax sample complexity of LFHT from works on the classical regime because the latter do not quantify the speed of convergence of the error terms as a function of the alphabet size. Specifically, let us examine [209, Theorem 1], which is a strengthening of the results of [90]. Paraphrasing, it states that for any fixed ratio $\alpha = n/m$ and pair of distributions $(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}})$, Gutman's test has type-II error bounded by $1/3$ when given samples from $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ as input, and type-I error bounded by $\exp(-\lambda n)$ given arbitrary input, where

$$\lambda = \mathrm{GJS}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \alpha) + \sqrt{\frac{V(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \alpha)}{n}}\Phi^{-1}(1/3) + \mathcal{O}\left(\frac{\log(n)}{n}\right) \tag{1}$$

as $n \to \infty$. Here V denotes the dispersion function defined in [209, Equation (9)] and Φ is the standard normal cdf. The crucial point we make here is that in (1) the dependence of the $\mathcal{O}(\log(n)/n)$ term on $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}$, and in particular their support size k and the ratio $\alpha = n/m$ is unspecified. Because of this, (1) and similar results cannot be used to derive minimax sample complexities as $\min\{n, m, k\} \to \infty$ jointly at possibly different rates.

This distinction between the fixed alphabet size setting studied in [90, 212, 209] and similar works, and our large alphabet setting was recognized by [102, 103, 123, 124] whose results are much closer to those of this paper. In [103] Huang and Meyn introduce the concept of "generalized error exponent" to deal with support sizes that grow superlinearly with sample size (referred to as the "sparse sample regime" by them) in the setting of uniformity testing.[3] In [102] they extend this idea to LFHT and say, quote,

> "In the classification problem, the classical error exponent analysis has been applied to the case of fixed alphabet in [212] and [90].... However, in the sparse sample problem, the classical error exponent concept is again not applicable, and thus a different scaling is needed."

Moving on to [123, 124], their authors study (2.1.2) with $n = m$ over the class of discrete distributions p with $\min_i p_i \asymp \max_i p_i \asymp 1/n^{\alpha}$, which they call *$\alpha$-large sources*. Disregarding the dependence on the TV-separation ϵ, effectively setting ϵ to a constant, they find that

[3] Uniformity testing is the problem of goodness-of-fit testing where the null is given by a uniform distribution.

achieving non-trivial minimax error is possible if and only if $\alpha \leq 2$, using in fact the same *difference of squared L^2-distances* test (2.1.1) that we study in this paper. Follow-up work [102] extends to the case $m \neq n$ and the class of distributions on alphabet $[k]$ with $\max_i p_i \lesssim 1/k$, we also cover this class under the name $\mathcal{P}_{\mathsf{Db}}$. In the regime of constant separation $\epsilon = \Theta(1)$ and $n, m \to \infty$ they show that LFHT with vanishing error is possible if and only if $k = o(\min(n^2, mn))$, thus discovering for the first time the *trade-off* between m and n.[4] Contrasting with our work, we are the first to characterize the full m, n, ϵ trade-off in the regime of constant probability of error, and we also consider three other classes of distributions, in addition to $\mathcal{P}_{\mathsf{Db}}$.

Another related problem is that of two-sample testing with unequal sample sizes, studied in [25, 64] for the class of discrete distributions $\mathcal{P}_{\mathsf{D}}$. In Section 2.3.1 we present reductions that show that our problem's sample complexity equals, up to constant factors, to that of two-sample testing in the case $m \geq n$. We emphasize that the distinction between $m \geq n$ and $m \leq n$ is necessary for this equivalence: in the latter case the sample complexities of the two problems are not the same. Moreover, our reduction doesn't help us solve classes other than $\mathcal{P}_{\mathsf{D}}$, as two-sample testing with unequal sample size exhibits a trade-off between n and m only in classes for which $n_{\mathsf{TS}} \neq n_{\mathsf{GoF}}$, see also the discussion at the end of Section 2.2.2.

The test (Scheffé) has been considered previously [58, 72, 144, 91, 143, 126, 94] and is also known as a "classification accuracy" test (CAT). Follow-up work [74] to the present paper shows that CATs are able to attain a (near-)minimax optimality in all settings studied here, and also achieve optimal dependence on the probability of error (in this paper we only consider a fixed error probability).

2.1.3 Contributions

Though the likelihood-free hypothesis testing problem (2.1.2) has previously appeared under various disguises and was studied in different regimes for the class of bounded discrete distributions, it omitted the key question of understanding the dependence of the sample complexity on the separation ϵ. Our work fully characterizes the dependence on the separation ϵ (Theorems 2.3.2 and 2.3.3). We discover the existence of a rather non-trivial trade-off between the m and n showing that in the likelihood-free setting statistical performance (m) can be traded for computational resources (n). Our results are shown for not just one but multiple distribution classes. In addition, we also demonstrate that LFHT naturally interpolates between its special cases corresponding to goodness-of-fit testing, two-sample testing and density-estimation. As a by-product we observe the relation $n_{\mathsf{GoF}}^2 \epsilon^2 \asymp n_{\mathsf{Est}}$ that holds over several classes of distributions and measures of separation, hinting at some universality property. On the technical side we provide a unified upper bound analysis for all regular classes we consider, and prove matching lower bounds using techniques of Tsybakov, Ingster and Valiant. Our upper bound analysis is inspired by Ingster [112, 110] whose L^2-distance testing approach, originally designed for goodness-of-fit in smooth-density

[4] The paper [102] contains implicitly other interesting results. For example, it appears that the constructive (upper bound) part of their proof if done carefully can also handle the case of variable $\epsilon \to 0$ in the regime $m, n \lesssim k$. Specifically, we believe they also show that for the minimax error $\delta \in (0, 1)$ LFHT is possible if $k \log(1/\delta)/\epsilon^4 \lesssim \min(n^2, nm)$. The lower bound appears to show LFHT is possible only if $k \log(1/\delta) \lesssim \min(n^2, nm)$. In addition they also apply the flattening technique, later re-discovered in [64].

classes, has been rediscovered in the discrete-alphabet world [123, 124, 81]. Compared to Ingster's work, the new ingredient needed in the discrete case is a "flattening" reduction [102, 64, 80], which we also utilize. Several minor extensions are also shown along the way, namely, robustness with respect to L^2-misspecification (Theorem 2.3.4) and characterization of n_{GoF} for the class of β-smooth densities with $\beta \leq 1$ under Hellinger separation (Theorem 2.3.6).

2.1.4 Structure

Section 2.2 defines the statistical problems and the classes of distributions that are studied throughout the paper, and discusses multiple tests for likelihood-free hypothesis testing. Section 2.3 contains our main results and the discussion linking to goodness-of-fit and two-sample testing, estimation and robustness. In Section 2.4 we provide sketch proofs for our results. Finally, in Section 2.5 we discuss possible future directions of research. The detailed proofs of Theorems 2.3.2 to 2.3.4 and 2.3.6 and all auxiliary results are included in the Appendix.

2.1.5 Notation

For $k \in \mathbb{N}$ we write $[k] =: \{1, 2, \dots, k\}$. For $x, y \in \mathbb{R}$ we write $x \wedge y =: \min\{x, y\}$, $x \vee y =: \max\{x, y\}$. We use the Bachmann–Landau notation $\Omega, \Theta, \mathcal{O}, o$ as usual and write $f \lesssim g$ for $f = \mathcal{O}(g)$ and $f \asymp g$ for $f = \Theta(g)$. For $c \in \mathbb{R}$ and $A \subseteq \mathbb{R}^2$ we write $cA =: \{(ca_1, ca_2) \in \mathbb{R}^2 : (a_1, a_2) \in A\}$. For two sets $A, B \subseteq \mathbb{R}^2$ we write $A \asymp B$ if there exists $c \in [1, \infty)$ with $\frac{1}{c}A \subseteq B \subseteq cA$. For two probability measures μ, ν dominated by η with densities p, q we define the following divergences: $\mathsf{TV}(\mu, \nu) =: \frac{1}{2}\int |p - q| \mathrm{d}\eta$, $\mathsf{H}(\mu, \nu) =: (\int (\sqrt{p} - \sqrt{q})^2 \mathrm{d}\eta)^{1/2}$, $\mathsf{KL}(\mu \| \nu) =: \int p \log(p/q) \mathrm{d}\eta$, $\chi^2(\mu \| \nu) =: \int ((p - q)^2/q) \mathrm{d}\eta$. Abusing notation, we sometimes write (p, q) as arguments instead of (μ, ν). We write $\| \cdot \|_p$ for the L^p and ℓ^p norms, where the base measure shall be clear from the context.

2.2 Sample complexity, non-parametric classes and tests

In the first two parts of this section we go over the technical background and definitions that are required to understand the rest of the paper, after which we give an exposition of multiple alternative approaches for our problem in Section 2.2.3.

2.2.1 Five fundamental problems in Statistics

Formally, we define a hypothesis as a set of probability measures. Given two hypotheses H_0 and H_1 on some space $\mathcal{X}$, we say that a function $\psi : \mathcal{X} \to \{0, 1\}$ successfully tests the two hypotheses against each other if

$$\max_{i=0,1} \sup_{P \in H_i} \mathbb{P}_{S \sim P}(\psi(S) \neq i) \leq 1/3. \tag{2.2.1}$$

Remark 13. *For our purposes, the constant* $1/3$ *above is unimportant and could be replaced by any number less than* $1/2$*. Throughout the paper we are interested in the asymptotic order of the sample complexity, and* $\Omega(\log(1/\delta))$*-way sample splitting followed by a majority vote decreases the overall error probability to* $\mathcal{O}(\delta)$ *of any successful tester, at the cost of inflating the sample complexity by a multiplicative* $\mathcal{O}(\log(1/\delta))$ *factor. Unfortunately, the resulting dependence on* δ *is sub-optimal except for binary hypothesis testing, see for example [16, Theorem 4.7]. Recent results for uniformity [63] and two-sample testing [62], and our follow-up work on LFHT [74] resolves the optimal dependence to be* $\sqrt{\log(1/\delta)}$ *or even* $\sqrt[3]{\log(1/\delta)}$ *in some regimes.*

Throughout this section let $\mathcal{P}$ be a class of probability distributions on $\mathcal{X}$. Suppose we observe independent samples $X \sim \mathbb{P}_{\mathsf{X}}^{\otimes n}$, $Y \sim \mathbb{P}_{\mathsf{Y}}^{\otimes n}$ and $Z \sim \mathbb{P}_{\mathsf{Z}}^{\otimes m}$ whose distributions $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Z}} \in \mathcal{P}$ are *unknown* to us. Finally, $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$ refer to distributions that are *known* to us. We now define five fundamental problems in statistics that we refer to throughout this paper.

Definition 1. ***Binary hypothesis testing*** *is the problem of testing*

$$H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_0 \qquad \textit{against} \qquad H_1 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_1 \tag{HT}$$

based on the sample X*. We use* $n_{\mathsf{HT}}(\epsilon, \mathcal{P})$ *to denote the* minimax sample complexity of binary hypothesis testing*, which is the smallest number such that for all* $n \geq n_{\mathsf{HT}}(\epsilon, \mathcal{P})$ *and all* $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$ *with* $\mathsf{TV}(\mathbb{P}_0, \mathbb{P}_1) \geq \epsilon$ *there exists a function* $\psi : \mathcal{X}^n \to \{0, 1\}$*, which given* X *as input successfully tests* H_0 *against* H_1 *in the sense of* (2.2.1)*.*

It is well known that the complexity of binary hypothesis testing is controlled by the Hellinger divergence.

Lemma 2.2.1. *For all* ϵ *and* $\mathcal{P}$ *with* $|\mathcal{P}| \geq 2$*, the relation*

$$n_{\mathsf{HT}}(\epsilon, \mathcal{P}) = \Theta\Big(\sup_{\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P} : \mathsf{TV}(\mathbb{P}_0, \mathbb{P}_1) \geq \epsilon} \mathsf{H}^{-2}(\mathbb{P}_0, \mathbb{P}_1) \Big)$$

holds, where the implied constant is universal.

Proof. We include the proof in Appendix A.4.1 for completeness. □

For all $\mathcal{P}$ considered in this paper $n_{\mathsf{HT}} = \Theta(1/\epsilon^2)$ holds. Therefore, going forward we usually refrain from the general notation n_{HT} and simply write $1/\epsilon^2$.

Definition 2. ***Goodness-of-fit testing*** *is the problem of testing*

$$H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_0 \qquad \textit{against} \qquad H_1 : \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_0) \geq \epsilon \textit{ and } \mathbb{P}_{\mathsf{X}} \in \mathcal{P} \tag{GoF}$$

based on the sample X*. We write* $n_{\mathsf{GoF}}(\epsilon, \mathcal{P})$ *for the* minimax sample complexity of goodness-of-fit testing*, which is the smallest value such that for all* $n \geq n_{\mathsf{GoF}}(\epsilon, \mathcal{P})$ *and* $\mathbb{P}_0 \in \mathcal{P}$ *there exists a function* $\psi : \mathcal{X}^n \to \{0, 1\}$*, which given* X *as input successfully tests* H_0 *against* H_1 *in the sense of* (2.2.1)*.*

Definition 3. ***Two-sample testing*** *is the problem of testing*

$$H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_{\mathsf{Z}} \text{ and } \mathbb{P}_{\mathsf{X}} \in \mathcal{P} \qquad \text{against} \qquad H_1 : \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Z}}) \geq \epsilon \text{ and } \mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Z}} \in \mathcal{P} \qquad (\mathsf{TS})$$

based on the samples X and Z. We write $\mathcal{R}_{\mathsf{TS}}(\epsilon, \mathcal{P})$ for the maximal subset of $\mathbb{R}^2$ such that for any $(n, m) \in \mathbb{N}^2$ for which there exists $(x, y) \in \mathcal{R}_{\mathsf{TS}}(\epsilon, \mathcal{P})$ with $(n, m) \geq (x, y)$ coordinate-wise, there also exists a function $\psi : \mathcal{X}^n \times \mathcal{X}^m \to \{0, 1\}$, which given X and Z as input successfully tests between H_0 and H_1 in the sense of (2.2.1). We will use the abbreviation $n_{\mathsf{TS}}(\epsilon, \mathcal{P}) = \min\{\ell \in \mathbb{N} : (\ell, \ell) \in \mathcal{R}_{\mathsf{TS}}(\epsilon, \mathcal{P})\}$ and refer to it as the minimax sample complexity of two-sample testing.

Definition 4. *The* minimax sample complexity of **estimation** *is the smallest value $n_{\mathsf{Est}}(\epsilon, \mathcal{P})$ such that for all $n \geq n_{\mathsf{Est}}(\epsilon, \mathcal{P})$ there exists an estimator $\widehat{\mathbb{P}}_{\mathsf{X}}$, which given X as input satisfies*

$$\mathbb{E}\mathsf{TV}(\widehat{\mathbb{P}}_{\mathsf{X}}, \mathbb{P}_{\mathsf{X}}) \leq \epsilon. \qquad (\mathsf{Est})$$

In order to simplify the presentation of our final definition, let us temporarily write $\mathcal{P}_\epsilon = \{(\mathbb{Q}_0, \mathbb{Q}_1) \in \mathcal{P}^2 : \mathsf{TV}(\mathbb{Q}_0, \mathbb{Q}_1) \geq \epsilon\}$. That is, $\mathcal{P}_\epsilon$ is the set of pairs of distributions in the class $\mathcal{P}$ which are ϵ separated in total variation.

Definition 5. ***Likelihood-free hypothesis testing*** *is the problem of testing*

$$H_0 : \mathbb{P}_{\mathsf{Z}} = \mathbb{P}_{\mathsf{X}} \text{ and } (\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \in \mathcal{P}_\epsilon \qquad \text{against} \qquad H_1 : \mathbb{P}_{\mathsf{Z}} = \mathbb{P}_{\mathsf{Y}} \text{ and } (\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \in \mathcal{P}_\epsilon \qquad (\mathsf{LF})$$

based on the samples X, Y and Z. Write $\mathcal{R}_{\mathsf{LF}}(\epsilon, \mathcal{P})$ for the maximal subset of $\mathbb{R}^2$ such that for any $(n, m) \in \mathbb{N}^2$ for which there exists $(x, y) \in \mathcal{R}_{\mathsf{LF}}(\epsilon, \mathcal{P})$ with $(n, m) \geq (x, y)$ coordinate-wise, there also exists a function $\psi : \mathcal{X}^n \times \mathcal{X}^n \times \mathcal{X}^m \to \{0, 1\}$, which given X, Y and Z as input successfully tests H_0 against H_1 in the sense of (2.2.1).

Requiring $\mathcal{R}_{\mathsf{TS}}(\epsilon, \mathcal{P})$ to be maximal is well defined, because for any $(n_0, m_0) \in \mathcal{R}_{\mathsf{TS}}(\epsilon, \mathcal{P})$ and $(n, m) \in \mathbb{N}^2$ with $(n_0, m_0) \leq (n, m)$ coordinate-wise, it must also hold that $(n, m) \in \mathcal{R}_{\mathsf{LF}}$, since ψ can simply disregard the extra samples. Clearly the same applies also to $\mathcal{R}_{\mathsf{LF}}(\epsilon, \mathcal{P})$.

Remark 14. *All five definitions above can be modified to measure separation with respect to an arbitrary function d instead of TV. We will write $n_{\mathsf{GoF}}(\epsilon, \mathsf{d}, \mathcal{P})$ et cetera for the corresponding values.*

2.2.2 Four classes of distributions

All of our definitions in the previous section assumed that we have some class of distributions $\mathcal{P}$ at hand. Below we introduce the classes that we study throughout the rest of the paper.

(i) **Smooth density.** Let $\mathcal{C}(\beta, d, C)$ denote the set of functions $f : [0, 1]^d \to \mathbb{R}$ that are $\underline{\beta} =: \lceil \beta - 1 \rceil$-times differentiable and satisfy

$$\|f\|_{\mathcal{C}_\beta} =: \max\left\{ \max_{0 \leq |\alpha| \leq \underline{\beta}} \|f^{(\alpha)}\|_\infty, \sup_{x \neq y \in [0,1]^d, |\alpha| = \underline{\beta}} \frac{|f^{(\alpha)}(x) - f^{(\alpha)}(y)|}{\|x - y\|_2^{\beta - \underline{\beta}}} \right\} \leq C,$$

where we write $|\alpha| = \sum_{i=1}^d \alpha_i$ for the multiindex $\alpha \in \mathbb{N}^d$ as usual. We further define $\mathcal{P}_{\mathsf{H}}(\beta, d, C)$ to be the class of distributions with Lebesgue-densities in $\mathcal{C}(\beta, d, C)$.

(ii) **Gaussian sequence model on the Sobolev ellipsoid.** Given $C > 0$ and a smoothness parameter $s > 0$, we define the Sobolev ellipsoid

$$\mathcal{E}(s, C) =: \Big\{\theta \in \mathbb{R}^{\mathbb{N}} : \sum_{j=1}^{\infty} j^{2s}\theta_j^2 \leq C\Big\}.$$

Our second distribution class is given by

$$\mathcal{P}_{\mathsf{G}}(s, C) =: \{\mu_\theta \ : \theta \in \mathcal{E}(s, C)\},$$

where $\mu_\theta = \otimes_{i=1}^{\infty}\mathcal{N}(\theta_i, 1)$. It is well known that this class models an s-smooth signal under Gaussian white noise, see for example [198, Section 1.7.1] for an exposition of this connection.

(iii)-(iv) **Distributions on a finite alphabet.** For $k \geq 2$, let

$$\mathcal{P}_{\mathsf{D}}(k) =: \{\text{all distributions on } \{1, 2, \ldots, k\}\}$$

denote the class of all discrete distributions, and set

$$\mathcal{P}_{\mathsf{Db}}(k, C) =: \{p \in \mathcal{P}_{\mathsf{D}}(k) : \|p\|_\infty \leq C/k\}$$

for all $C > 1$. In other words, $\mathcal{P}_{\mathsf{Db}}$ are those distributions with support in $\{1, 2, \ldots, k\}$ that are bounded by a constant multiple of the uniform distribution.

Note that depending on the choice of C some of the above distribution classes may be empty. To avoid such issues, throughout the rest of paper we implicitly operate under the following assumption.

Assumption 1. *We always assume that $C > 1$ when referring to $\mathcal{P}_{\mathsf{H}}(\beta, d, C)$ and $\mathcal{P}_{\mathsf{Db}}(k, C)$.*

As we shall see in Section 2.3.2 when discussing our results, the behaviour of $\mathcal{P}_{\mathsf{D}}$ is qualitatively different from the other three classes introduced above. Consequently, we will sometimes refer to $\mathcal{P}_{\mathsf{Db}}$ as the "regular discrete" class, and we will see that its minimax sample complexities are similar to $\mathcal{P}_{\mathsf{H}}$ and $\mathcal{P}_{\mathsf{G}}$ but different from $\mathcal{P}_{\mathsf{D}}$. More generally we will call the classes $\mathcal{P}_{\mathsf{H}}, \mathcal{P}_{\mathsf{G}}, \mathcal{P}_{\mathsf{Db}}$ "regular", characterized by the fact that $n_{\mathsf{GoF}} \asymp n_{\mathsf{TS}}$, or equivalently, by the fact that $\mathcal{R}_{\mathsf{TS}} \asymp \{(n, m) : \min\{n, m\} \geq n_{\mathsf{TS}}\}$.

2.2.3 Tests for LFHT

We start this section by reintroducing the difference of L^2-distances statistic that our results are based on, and which we've already seen in (2.1.1). Then, in Section 2.2.3 we mention some natural alternative approaches to the problem, which we however do not study further. Therefore, the reader that wishes to proceed to our results without delay may safely skip over Section 2.2.3.

Ingster's L^2-distance test

For simplicity we focus on the case of discrete distributions. This case is more general than may first appear: for example in the case of smooth densities on $[0,1]^d$ one can simply take a regular grid (whose resolution is determined by the smoothness of the densities) and count the number of datapoints falling in each cell. Let $\widehat{p}_{\mathsf{X}}, \widehat{p}_{\mathsf{Y}}, \widehat{p}_{\mathsf{Z}}$ denote the empirical probability mass functions of the finitely supported distributions $\widehat{\mathbb{P}}_{\mathsf{X}}, \widehat{\mathbb{P}}_{\mathsf{Y}}, \widehat{\mathbb{P}}_{\mathsf{Z}}$. The test proceeds via the comparison

$$\|\widehat{p}_{\mathsf{X}} - \widehat{p}_{\mathsf{Z}}\|_2 \lesseqgtr \|\widehat{p}_{\mathsf{Y}} - \widehat{p}_{\mathsf{Z}}\|_2. \tag{2.2.2}$$

Squaring both sides and rearranging, we arrive at the form

$$\frac{1}{m}\sum_{i=1}^{m}(\widehat{p}_{\mathsf{Y}}(Z_i) - \widehat{p}_{\mathsf{X}}(Z_i)) \lesseqgtr \gamma,$$

where $\gamma = (\|\widehat{p}_{\mathsf{Y}}\|^2 - \|\widehat{p}_{\mathsf{X}}\|^2)/2$. As mentioned in the introduction, variants of this L^2-distance based test have been invented and re-invented multiple times for goodness-of-fit [110, 81] and two-sample testing [21, 10]. The exact statistic (2.2.2) with application to $\mathcal{P}_{\mathsf{Db}}$ has appeared in [123, 124], and Huang and Meyn [102] proposed an ingenious improvement restricting attention exclusively to bins whose counts are one of $(2,0),(1,1),(0,2)$ for the samples (X,Z) or (Y,Z). We attribute (2.2.2) to Ingster because his work on goodness-of-fit testing for smooth densities is the first occurence of the idea of comparing empirical L^2 norms, but we note that [123] and [81] arrive at this influential idea apparently independently.

We emphasize the following subtlety. Let us rewrite (2.2.2) as

$$\|\widehat{p}_{\mathsf{X}} - \widehat{p}_{\mathsf{Z}}\|_2^2 - \|\widehat{p}_{\mathsf{Y}} - \widehat{p}_{\mathsf{Z}}\|_2^2 \lesseqgtr 0\,. \tag{2.2.3}$$

As we shall see from our proofs, this difference results in an optimal test for the full range of possible values of n and m for $\mathcal{P}_{\mathsf{Db}}$. However, this does not mean that each term by itself is a meaningful estimate of the corresponding distance: rejecting the null by thresholding just $\|\widehat{p}_{\mathsf{X}} - \widehat{p}_{\mathsf{Z}}\|_2^2$ would not work. Indeed, the variance of $\|p_{\mathsf{X}} - \widehat{p}_{\mathsf{Z}}\|_2^2$ is so large that one needs $m \gtrsim n_{\mathsf{GoF}} \gg 1/\epsilon^2$ observations to obtain a reliable estimate of $\|p_{\mathsf{X}} - p_{\mathsf{Z}}\|_2^2$. The "magic" of the L^2-difference test is that the two terms in (2.2.3) separately have high variance, and thus are not good estimators of their means, but their difference cancels the high-variance terms.

Remark 15. *While testing* (LF)*, practitioners are usually interested in obtaining a p-value, rather than purely a decision whether to reject the null hypothesis. For this we propose the following scheme. Let $\sigma_1, \ldots, \sigma_P$ be i.i.d. uniformly random permutations on $n+m$ elements. Let $\widehat{T} = \|\widehat{p}_{\mathsf{X}} - \widehat{p}_{\mathsf{Z}}\|_2^2 - \|\widehat{p}_{\mathsf{Y}} - \widehat{p}_{\mathsf{Z}}\|_2^2$ be our statistic, and write $\widehat{T}_i$ for the statistic $\widehat{T}$ evaluated on the permuted dataset where $\{X_1, \ldots, X_n, Z_1, \ldots, Z_m\}$ are shuffled according to σ_i. Under the null the random variables $\widehat{T}, \widehat{T}_1, \ldots, \widehat{T}_P$ are exchangeable, thus reporting the empirical upper quantile of $\widehat{T}$ in this sample yields an estimate of the p-value. Studying the variance of this estimate or the power of the test that rejects when the estimated p-value is less than some threshold, is beyond the scope of this work.*

Alternative tests for LFHT

In this section we discuss a variety of alternative tests that may be considered for (LF) instead of (2.2.3). These are included only to provide additional context for our problem, and the reader may safely skip it and proceed to our results in Section 2.3. The approaches we consider are

(i) Scheffé's test,

(ii) Likelihood-free Neyman-Pearson test and

(iii) Huber's and Birgé's robust tests.

The tests (*i-ii*) are based on the idea of using the simulated samples to learn a set or a function that separates $\mathbb{P}_{\mathsf{X}}$ from $\mathbb{P}_{\mathsf{Y}}$. The test (*iii*) and (2.2.3) use the simulated samples to obtain density estimates of $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}$ directly. All of them, however, are of the form

$$\sum_{i=1}^{m} s(Z_i) \lesseqgtr 0 \tag{2.2.4}$$

with only the function s varying.

Variants of *Scheffé's test* using machine-learning enabled classifiers are the subject of current research in two-sample testing [144, 143, 91, 126, 94] and are used in practice for LFI specifically in high energy physics, see also our discussion of the Higgs boson discovery in Section 2.1. Thus, understanding the performance of Scheffé's test in the context of (LF) is of great practical importance. Suppose that using the simulated samples we train a probabilistic classifier $C : \mathcal{X} \to [0, 1]$ on the labeled data $\cup_{i=1}^{n}\{(X_i, 0), (Y_i, 1)\}$. The specific form of the classifier here is arbitrary and can be anything from logistic regression to a deep neural network. Given thresholds $t, \gamma \in [0, 1]$ chosen to satisfy our risk appetite for type-I vs type-II errors, Scheffé's test proceeds via the comparison

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{C(Z_i) \geq t\} \lesseqgtr \gamma. \tag{2.2.5}$$

We see that (2.2.5) is of the form (2.2.4) with $s(z) = (\mathbb{1}\{C(z) \geq t\} - \gamma)/m$. The follow-up work [74] studies the performance of Scheffé's test in great detail, finding that it is (near-)minimax optimal in all cases considered in this paper. It is found that the optimal classifier C must be trained *not* purely to minimize misclassification error, but rather must also keep the variance of its output small.

If the distributions $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}$ are fully known, then the likelihood-ratio test corresponds to

$$\sum_{i=1}^{m} s_{\mathsf{NP}}(Z_i) \lesseqgtr \gamma \qquad s_{\mathsf{NP}}(z) = \log\left(\frac{\mathrm{d}\mathbb{P}_{\mathsf{X}}}{\mathrm{d}\mathbb{P}_{\mathsf{Y}}}(z)\right), \tag{2.2.6}$$

where γ is again chosen to satisfy our type-I vs type-II error trade-off preferences. It is well known that the above procedure is optimal due to the Neyman-Pearson lemma. Recall that in our setting $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}$ are known only up to i.i.d. samples, and therefore it seems natural to

try to estimate s_{NP} from samples. It is not hard to see that s_{NP} minimizes the population *cross-entropy/logistic loss*, that is

$$s_{\mathsf{NP}} = \arg\min_s \mathbb{E}_{z\sim\mathbb{P}_{\mathsf{X}}}[\ell(s(z),1)] + \mathbb{E}_{z\sim\mathbb{P}_{\mathsf{Y}}}[\ell(s(z),0)]\,,$$

where $\ell(s,y) = \log(1+e^s) - ys$. In practice, the majority of today's classifiers are obtained by running some form of gradient descent on the problem

$$\widehat{s} = \arg\min_{s\in\mathcal{G}} \mathbb{E}_{z\sim\widehat{\mathbb{P}}_{\mathsf{X}}}[\ell(s(z),1)] + \mathbb{E}_{z\sim\widehat{\mathbb{P}}_{\mathsf{Y}}}[\ell(s(z),0)]\,,$$

where $\mathcal{G}$ is, say, a parametric class of neural networks and $\widehat{\mathbb{P}}_{\mathsf{X}}, \widehat{\mathbb{P}}_{\mathsf{Y}}$ are empirical distributions. Given such an estimate $\widehat{s}$, we can replace the unknown s_{NP} in (2.2.6) by $\widehat{s}$ to obtain the *likelihood-free Neyman-Pearson test*. For recent work on this approach in LFI see for example [53]. Studying properties of this test is outside the scope of this paper.

The final approach is based on the idea of *robust testing*, first proposed by Huber [104, 105]. Huber's seminal result implies that if one has approximately correct distributions $\widehat{\mathbb{P}}_{\mathsf{X}}, \widehat{\mathbb{P}}_{\mathsf{Y}}$ satisfying

$$\max\left\{\mathsf{TV}(\widehat{\mathbb{P}}_{\mathsf{X}}, \mathbb{P}_{\mathsf{X}}), \mathsf{TV}(\widehat{\mathbb{P}}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Y}})\right\} \le \epsilon/3 \qquad \text{and} \qquad \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \ge \epsilon,$$

then for some $c_1 < c_2$ the test

$$\sum_{i=1}^{m} s_{\mathsf{H}}(Z_i) \lessgtr 0 \qquad \text{where} \qquad s_{\mathsf{H}}(z) = \min\left\{\max\left\{c_1, \log\left(\frac{\mathrm{d}\widehat{\mathbb{P}}_{\mathsf{X}}}{\mathrm{d}\widehat{\mathbb{P}}_{\mathsf{Y}}}(z)\right)\right\}, c_2\right\}$$

has type-I and type-II error bounded by $\exp(-\Omega(m\epsilon^2))$, and is in fact minimax optimal for all sample sizes analogously to the likelihood-ratio test in the case of binary hypothesis testing. From the above formula we can see that Scheffé's test can be interpreted as an approximation of the maximally robust Huber's test. Let $\widehat{\mathcal{L}}(z) = (\mathrm{d}\widehat{\mathbb{P}}_{\mathsf{Y}}/\mathrm{d}\widehat{\mathbb{P}}_{\mathsf{X}})(z)$ denote the likelihood-ratio of the estimates. The values of c_1, c_2 are given as the solution to

$$\epsilon/3 = \mathbb{E}_{z\sim\widehat{\mathbb{P}}_{\mathsf{X}}}\left[\mathbb{1}\left\{\widehat{\mathcal{L}}(z) \le c_1\right\}\frac{c_1 - \widehat{\mathcal{L}}(z)}{1+c_1}\right] = \mathbb{E}_{z\sim\widehat{\mathbb{P}}_{\mathsf{Y}}}\left[\mathbb{1}\left\{\widehat{\mathcal{L}}(z) \ge c_2\right\}\frac{\widehat{\mathcal{L}}(z) - c_2}{1+c_2}\right],$$

which can be easily approximated to high accuracy given samples from $\widehat{\mathbb{P}}_{\mathsf{X}}, \widehat{\mathbb{P}}_{\mathsf{Y}}$. This suggests both a theoretical construction, since $\widehat{\mathbb{P}}_{\mathsf{X}}, \widehat{\mathbb{P}}_{\mathsf{Y}}$ can be obtained with high probability from simulation samples via the general estimator of Yatracos [206], and a practical rule: instead of the possibly brittle likelihood-free Neyman-Pearson test (ii), one should try clamping the estimated log-likelihood ratio from above and below.

Similar results hold due to Birgé [28, 27] in the case when distance is measured by Hellinger divergence:

$$\max\left\{\mathsf{H}(\widehat{\mathbb{P}}_{\mathsf{X}}, \mathbb{P}_{\mathsf{X}}), \mathsf{H}(\widehat{\mathbb{P}}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Y}})\right\} \le \epsilon/3 \qquad \text{and} \qquad \mathsf{H}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \ge \epsilon.$$

For ease of notation, let $\widehat{p}_{\mathsf{X}}, \widehat{p}_{\mathsf{Y}}$ denote the densities of $\widehat{\mathbb{P}}_{\mathsf{X}}, \widehat{\mathbb{P}}_{\mathsf{Y}}$ with respect to some base measure μ. Regarding $\sqrt{\widehat{p}_{\mathsf{X}}}$ and $\sqrt{\widehat{p}_{\mathsf{Y}}}$ as unit vectors of the Hilbert space $L^2(\mu)$, let $\gamma : [0,1] \to L^2(\mu)$

be the constant speed geodesic on the unit sphere of $L^2(\mu)$ with $\gamma(0) = \sqrt{\widehat{p}_{\mathsf{X}}}$ and $\gamma(1) = \sqrt{\widehat{p}_{\mathsf{Y}}}$. It is easily checked that each γ_t is positive, and Birgé showed that the test

$$\sum_{i=1}^{m} \log\left(\frac{\gamma_{1/3}^2}{\gamma_{2/3}^2}(Z_i)\right) \lesseqgtr 0$$

has both type-I and type-II errors bounded by $\exp(-\Omega(m\epsilon^2))$. For an exposition of this result see also [79, Theorem 7.1.2]

2.3 Results

In this section we describe our results on the sample complexity of likelihood-free hypothesis testing.

2.3.1 General reductions

In this first section, we give reductions that hold in great generality and show the relationship of our problem with other classical testing and estimation problems that were introduced in Section 2.2.1. The result below holds for a generic class $\mathcal{P}$ of distributions and a generic measure of separation d, see also Remark 14.

Proposition 2.3.1. *Let $\mathcal{P}$ be a generic family of distributions and $\mathsf{d} : \mathcal{P}^2 \to \mathbb{R}$ be any function used to measure separation. There exists a universal constant $c > 0$ such that for $n, m \in \mathbb{N}$ the following implications hold.*

$$(n,m) \in \mathcal{R}_{\mathsf{LF}} \implies m \geq n_{\mathsf{HT}}, \tag{2.3.1}$$
$$(n,m) \in \mathcal{R}_{\mathsf{TS}} \implies n \wedge m \geq n_{\mathsf{GoF}} \tag{2.3.2}$$
$$(n,m) \in \mathcal{R}_{\mathsf{LF}} \implies cn \geq n_{\mathsf{GoF}}, \tag{2.3.3}$$
$$(n,m) \in \mathcal{R}_{\mathsf{TS}} \implies (n,m) \in \mathcal{R}_{\mathsf{LF}}, \tag{2.3.4}$$
$$m \geq n \text{ and } (n,m) \in \mathcal{R}_{\mathsf{LF}} \implies (cn, cm) \in \mathcal{R}_{\mathsf{TS}}, \tag{2.3.5}$$

where we omit the argument $(\epsilon, \mathsf{d}, \mathcal{P})$ throughout for simplicity. In particular,

$$\mathbb{N}^2_{n \leq m} \cap \mathcal{R}_{\mathsf{LF}} \asymp \mathbb{N}^2_{n \leq m} \cap \mathcal{R}_{\mathsf{TS}}, \tag{2.3.6}$$

where $\mathbb{N}^2_{n \leq m} = \{(n,m) \in \mathbb{N}^2 : n \leq m\}$.

Proof. In what follows, let $\Psi_{\mathsf{LF}}, \Psi_{\mathsf{TS}}$ be minimax optimal tests for (LF) and (TS) respectively. Throughout the proof we omit the arguments $(\epsilon, \mathsf{d}, \mathcal{P})$ for notational simplicty.

Reducing hypothesis testing to (LF) Suppose $(n,m) \in \mathcal{R}_{\mathsf{LF}}$. Let $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$ be given with $\mathsf{d}(\mathbb{P}_0, \mathbb{P}_1) \geq \epsilon$ and suppose Z is an i.i.d. sample with m observations. We wish to test the hypothesis $H_0 : Z_i \sim \mathbb{P}_0$ against $H_1 : Z_i \sim \mathbb{P}_1$. To this end generate n i.i.d. observations X, Y from $\mathbb{P}_0, \mathbb{P}_1$ respectively, and simply output $\Psi_{\mathsf{LF}}(X, Y, Z)$. This shows that if $(n,m) \in \mathcal{R}_{\mathsf{LF}}$ then $m \geq n_{\mathsf{HT}}$ and concludes the proof of (2.3.1).

Reducing goodness-of-fit testing to two-sample testing Suppose $(n, m) \in \mathcal{R}_{\mathsf{TS}}$. Then obviously $(n \wedge m, \infty) \in \mathcal{R}_{\mathsf{TS}}$. However, two-sample testing with sample sizes $n \wedge m, \infty$ is equivalent to goodness-of-fit testing with a sample size of $n \wedge m$. Therefore, $n \wedge m \geq n_{\mathsf{GoF}}$ must hold, concluding the proof of (2.3.2).

Reducing goodness-of-fit testing to (LF) Suppose $(n, m) \in \mathcal{R}_{\mathsf{LF}}$ with $m \leq n$. Let a distribution $\mathbb{P}_0 \in \mathcal{P}$ be given as well as an i.i.d. sample X of size cn with unknown distribution $\mathbb{P}_{\mathsf{X}}$, where $c \in \mathbb{N}$ is a large integer. We want to test $H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_0$ against $H_1 : \mathbb{P}_{\mathsf{X}} \in \mathcal{P}, \mathsf{d}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_0) \geq \epsilon$. Generate $c \times 2$ i.i.d. samples $Y^{(i)}, Z^{(i)}$ for $i = 1, \ldots, c$ of size n, m respectively, all from $\mathbb{P}_0$. Split the sample X into c batches $X^{(i)}, i = 1, \ldots, c$ of size n each and form the variables

$$A_i = \Psi_{\mathsf{LF}}(X^{(i)}, Y^{(i)}, Z^{(i)}) - \Psi_{\mathsf{LF}}(X^{(i)}, Y^{(i)}, X^{(i+1)}_{1:m})$$

for $i = 1, 3, \ldots, 2\lfloor c/2 \rfloor - 1$, where $X^{(i)}_{1:m}$ denotes the first m observations in the batch $X^{(i)}$. Note that the A_i are i.i.d. and bounded random variables. Under the null hypothesis we have $\mathbb{E}A_i = 0$, while under the alternative they have mean $\mathbb{E}A_i \geq 1/3$ (since Ψ_{LF} is a successful tester in the sense of (2.2.1)). Therefore, a constant number $c/2$ observations suffice to decide whether $\mathbb{P}_{\mathsf{X}} = \mathbb{P}_0$ or not. In particular, $cn \geq n_{\mathsf{GoF}}$ which concludes the proof of (2.3.3) for the case $m \leq n$. The case $n \leq m$ follows from (2.3.5) and (2.3.2).

Reducing (LF) to two-sample testing Suppose $(n, m) \in \mathcal{R}_{\mathsf{TS}}$. Let three samples X, Y, Z be given, of sizes a, a, b from the unknown distributions $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Z}}$ respectively, where $\{a, b\} = \{n, m\}$. We want to test the hypothesis $H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_{\mathsf{Z}}$ against $H_1 : \mathbb{P}_{\mathsf{Y}} = \mathbb{P}_{\mathsf{Z}}$, where $\mathsf{d}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \geq \epsilon$ under both. Then, the test

$$\widetilde{\Psi}_{\mathsf{LF}}(X, Y, Z) =: \Psi_{\mathsf{TS}}(X, Z)$$

shows that $(n, m), (m, n) \in \mathcal{R}_{\mathsf{LF}}$ and concludes the proof of (2.3.4).

Reducing two-sample testing to (LF) Suppose $(n, m) \in \mathcal{R}_{\mathsf{LF}}$ where $m \geq n$. Let two samples X, Y be given, from the unknown distributions $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P}$ and of sample size cn, cm respectively, where $c \in \mathbb{N}$ is a large integer. We wish to test the hypothesis $H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_{\mathsf{Y}}$ against $H_1 : \mathsf{d}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \geq \epsilon$. Split the samples X, Y into $2 \times c$ batches $X^{(i)}, Y^{(i)}, i = 1, \ldots, c$ of sizes n, m respectively, and form the variables

$$A_i = \Psi_{\mathsf{LF}}(X^{(i)}, Y^{(i)}_{1:n}, Y^{(i+1)}) - \Psi_{\mathsf{LF}}(Y^{(i)}_{1:n}, X^{(i)}, Y^{(i+1)})$$

for $i = 1, 3, \ldots, 2\lfloor c/2 \rfloor - 1$, where $Y^{(i)}_{1:n}$ denotes the first n observations in the batch $Y^{(i)}$. The variables A_i are i.i.d. and bounded. Under the null hypothesis we have $\mathbb{E}A_i = 0$ while under the alternative $\mathbb{E}A_i \geq 1/3$ holds. Therefore a constant number $c/2$ observations suffice to decide whether $\mathbb{P}_{\mathsf{X}} = \mathbb{P}_{\mathsf{Y}}$ or not. In particular, $(cn, cm) \in \mathcal{R}_{\mathsf{TS}}$ which concludes the proof of (2.3.5).

Equivalence between two-sample testing and (LF) Equation (2.3.6) follows immediately from (2.3.5) and (2.3.4). □

Equation (2.3.6) tells us that the problems of likelihood-free hypothesis testing and two-sample testing are equivalent, *but only for* $m \geq n$, that is, when we have more real data than simulated data. We will see in the next section, and on Figure 2.1 visually, that this distinction is necessary.

2.3.2 Sample complexity of likelihood-free hypothesis testing

In this section we present our results on the sample complexity of (LF) for the specific classes $\mathcal{P}$ that were introduced in Section 2.2.1, with separation measured by TV. In all results below the parameters β, s, d, C are regarded as constants, we only care about the dependence on the separation ϵ and the alphabet size k (in the case of $\mathcal{P}_{\mathsf{D}}, \mathcal{P}_{\mathsf{Db}}$). Where convenient we omit the arguments of $n_{\mathsf{GoF}}, n_{\mathsf{TS}}, \mathcal{R}_{\mathsf{TS}}, n_{\mathsf{Est}}, \mathcal{R}_{\mathsf{LF}}$ to ease notation, whose value should be clear from the context.

Theorem 2.3.2. *Under* TV*-separation, for each choice* $\mathcal{P} \in \{\mathcal{P}_{\mathsf{H}}, \mathcal{P}_{\mathsf{G}}, \mathcal{P}_{\mathsf{Db}}\}$*, we have*

$$\mathcal{R}_{\mathsf{LF}} \asymp \Big\{(n,m) \,:\, m \geq 1/\epsilon^2,\, n \geq n_{\mathsf{GoF}},\, mn \geq n_{\mathsf{GoF}}^2\Big\},$$

where the implied constants do not depend on k *(in the case of* $\mathcal{P}_{\mathsf{Db}}$*) or* ϵ*.*

For each class $\mathcal{P}$ in Theorem 2.3.2, the entire region $\mathcal{R}_{\mathsf{LF}}$ (within universal constant) is attained by a suitable modification of Ingster's L^2-distance test from Section 2.2.3. The region $\mathcal{R}_{\mathsf{LF}}$ is visualized on Figure 2.1 on a log-log scale, with each corner point $\{\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}\}$ having a special interpretation, as per the reductions presented in Proposition 2.3.1. The point A corresponds to binary hypothesis testing and D can be reduced to goodness-of-fit testing. Similarly, B and C can be reduced to the well-known problems of estimation followed by robust hypothesis testing and two-sample testing respectively. In other words, (LF) allows us to naturally interpolate between multiple statistical problems. Finally, we make an interesting observation: since the product of n and m remains constant on the line segment $[\mathsf{B}, \mathsf{C}]$ on the left plot of Figure 2.1, it follows that

$$n_{\mathsf{Est}}(\epsilon, \mathcal{P}) \asymp n_{\mathsf{GoF}}^2(\epsilon, \mathcal{P})\, \epsilon^2 \tag{2.3.7}$$

for each class $\mathcal{P}$ treated in Theorem 2.3.2. This relation between the sample complexity of estimation and goodness-of-fit testing has not been observed before to our knowledge, and understanding the scope of validity of this relationship is an exciting future direction.[5]

Turning to our results on $\mathcal{P}_{\mathsf{D}}$ the picture is less straightforward. As first identified in [20] and fully resolved in [39], the sample complexity of two-sample testing undergoes a phase transition when $k \gtrsim 1/\epsilon^4$. This phase transition appears also in likelihood-free hypothesis testing.

Theorem 2.3.3. *Let* $\alpha = \max\big\{1, \min\big\{\frac{k}{n}, \frac{k}{m}\big\}\big\}$*. Then*

$$\mathcal{R}_{\mathsf{LF}}(\epsilon, \mathcal{P}_{\mathsf{D}}(k)) \asymp_{\log(k)} \Big\{(n,m) \,:\, m \geq 1/\epsilon^2,\, n \geq n_{\mathsf{GoF}}(\epsilon, \mathcal{P}_{\mathsf{D}}(k))\cdot\sqrt{\alpha},\, mn \geq n_{\mathsf{GoF}}(\epsilon, \mathcal{P}_{\mathsf{D}}(k))^2\cdot\alpha\Big\},$$

where the equivalence is up to a logarithmic factor in the alphabet size k*.*

The $\log k$ factor in our analysis originates from a union bound, and it is possible that it may be removed. It follows from follow up work [74] and past results on two-sample testing

[5] *Added in print:* for example in [169] it is demonstrated that for the Gaussian sequence model (see definition (*ii*) in Section 2.2.2) with the Sobolev ellipsoid replaced by the set $\Theta = \{\theta \in \ell^2 : \sum_{i=1}^{\infty} i|\theta_i| \leq 1\}$, it holds that $n_{\mathsf{Est}} \ll n_{\mathsf{GoF}}^2/\epsilon^2$.

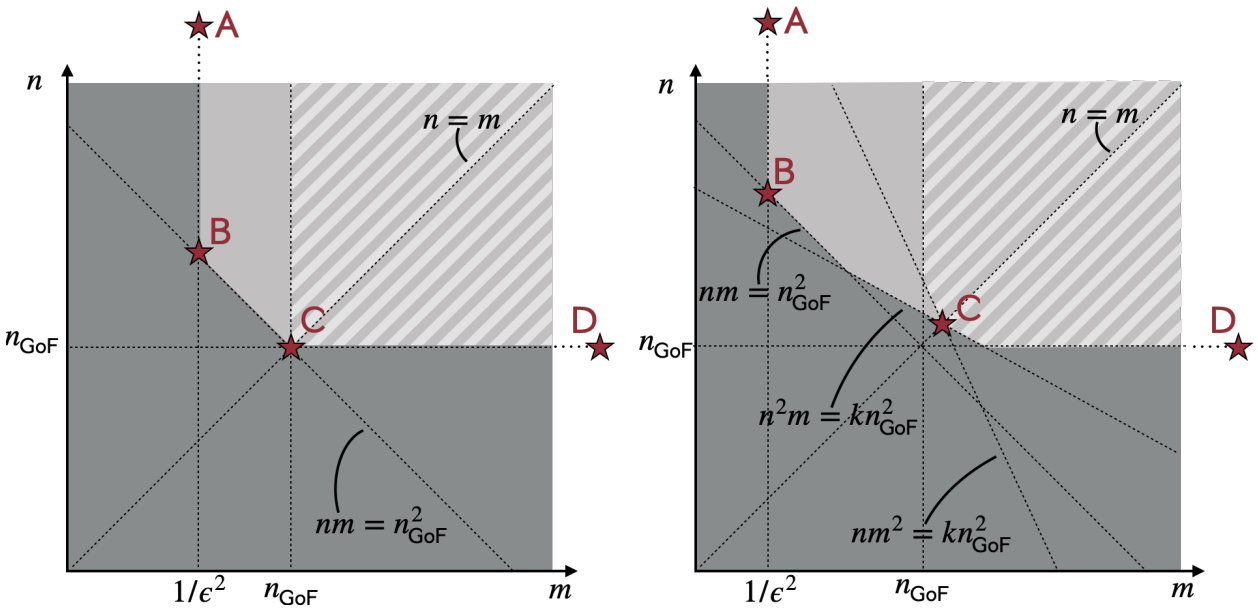


Figure 2.1: Light and dark gray show $\mathcal{R}_{\mathsf{LF}}$ and its complement resp. on log scale; the striped region depicts $\mathcal{R}_{\mathsf{TS}} \subsetneq \mathcal{R}_{\mathsf{LF}}$. Left plot is valid for $\mathcal{P} \in \{\mathcal{P}_{\mathsf{H}}, \mathcal{P}_{\mathsf{G}}, \mathcal{P}_{\mathsf{Db}}\}$ for all settings of ϵ, k. For $\mathcal{P}_{\mathsf{D}}$ the left plot applies when $k \lesssim \epsilon^{-4}$ and the right plot otherwise.

[62] that the $\log(k)$ factor can be removed in all regimes, thus fully characterizing the sample complexity of (LF), but using a different test from ours.

Table 2.1 summarizes previously known tight results for the values of $n_{\mathsf{GoF}}, n_{\mathsf{TS}}, \mathcal{R}_{\mathsf{TS}}$ and n_{Est}. The fact that $n_{\mathsf{HT}} = \Theta(1/\epsilon^2)$ for reasonable classes is classical, see Lemma 2.2.1. The study of goodness-of-fit testing within a minimax framework was pioneered by Ingster [112, 110] for $\mathcal{P}_{\mathsf{H}}, \mathcal{P}_{\mathsf{G}}$, and independently studied by the computer science community [81, 199] for $\mathcal{P}_{\mathsf{D}}, \mathcal{P}_{\mathsf{Db}}$ under the name *identity testing*. Two-sample testing (a.k.a. *closeness testing*) was solved in [39] for $\mathcal{P}_{\mathsf{D}}$ (with the optimal result for $\mathcal{P}_{\mathsf{Db}}$ implicit) and [110, 10, 139] consider $\mathcal{P}_{\mathsf{H}}$. The study of the rate of estimation n_{Est} is older, see [108, 198, 119, 79] and references for $\mathcal{P}_{\mathsf{H}}, \mathcal{P}_{\mathsf{G}}$ and [35] for $\mathcal{P}_{\mathsf{D}}, \mathcal{P}_{\mathsf{Db}}$.

Table 2.1: Prior results on testing and estimation

	n_{HT}	n_{GoF}	$\mathcal{R}_{\mathsf{TS}}$	n_{Est}
$\mathcal{P}_{\mathsf{G}}$	$1/\epsilon^2$	$1/\epsilon^{(2s+1/2)/s}$	$n \wedge m \geq n_{\mathsf{GoF}}$	$\epsilon^2\, n_{\mathsf{GoF}}^2$
$\mathcal{P}_{\mathsf{H}}$	$1/\epsilon^2$	$1/\epsilon^{(2\beta+d/2)/\beta}$	$n \wedge m \geq n_{\mathsf{GoF}}$	$\epsilon^2\, n_{\mathsf{GoF}}^2$
$\mathcal{P}_{\mathsf{Db}}$	$1/\epsilon^2$	$\sqrt{k}/\epsilon^2$	$n \wedge m \geq n_{\mathsf{GoF}}$	$\epsilon^2\, n_{\mathsf{GoF}}^2$
$\mathcal{P}_{\mathsf{D}}$	$1/\epsilon^2$	$\sqrt{k}/\epsilon^2$	$n \vee m \geq \frac{\sqrt{k}}{\epsilon^2} \vee \frac{k^{2/3}}{\epsilon^{4/3}} \asymp n_{\mathsf{TS}},\ n \wedge m \geq n_{\mathsf{GoF}}\sqrt{\alpha}$	$\epsilon^2\, n_{\mathsf{GoF}}^2$

2.3.3 L^2-robust likelihood-free hypothesis testing

Even before seeing Theorems 2.3.2 and 2.3.3 one might guess that estimation in TV followed by a robust hypothesis test should work whenever $m \gtrsim 1/\epsilon^2$ and $n \geq n_{\mathsf{Est}}(c\epsilon)$ for a small enough constant c. This strategy does indeed work, which can be deduced from the work of Huber and Birgé [104, 27] for total variation and Hellinger separation respectively, see also Section 2.2.3 for a brief discussion of these robust tests. In other words, we have the informal theorem

$$\text{if separation is measured by } \mathsf{TV} \text{ or } \mathsf{H}, \text{ then } (n \geq n_{\mathsf{Est}} \text{ and } m \geq n_{\mathsf{HT}}) \implies (cn, cm) \in \mathcal{R}_{\mathsf{LF}}.$$

In the case of total variation separation, in fact an even simpler approach succeeds: if $\widehat{p}_{\mathsf{X}}$ and $\widehat{p}_{\mathsf{Y}}$ are minimax optimal density estimators with respect to TV, then Scheffé's test using the classifier $C(x) = \mathbb{1}\{\widehat{p}_{\mathsf{Y}}(x) \geq \widehat{p}_{\mathsf{X}}(x)\}$ can be shown to achieve the optimal sample complexity by Chebyshev's inequality.

The upshot of these observations is that they provide a solution to (LF) that is robust to model misspecification, specifically at the corner point B on Figure 2.1. This naturally leads us to the question of robust likelihood-free hypothesis testing: can we construct robust tests for the full m vs n trade-off?

As before, suppose we observe samples X, Y, Z of size n, n, m from distributions belonging to the class $\mathcal{P}$ with densities f, g, h with respect to some base measure μ. Given any $u \in \mathcal{P}$, let $\mathsf{B}_u(\epsilon, \mathcal{P}) \subseteq \mathcal{P}$ denote a region around u against which we wish to be robust. Recall the notation $\mathcal{P}_\epsilon = \{(\mathbb{Q}_0, \mathbb{Q}_1) \in \mathcal{P}^2 : \mathsf{TV}(\mathbb{Q}_0, \mathbb{Q}_1) \geq \epsilon\}$ from Definition 5. We compare the hypotheses

$$H_0 : h \in \mathsf{B}_f(\epsilon, \mathcal{P}), (f, g) \in \mathcal{P}_\epsilon \qquad \text{versus} \qquad H_1 : h \in \mathsf{B}_g(\epsilon, \mathcal{P}), (f, g) \in \mathcal{P}_\epsilon, \tag{rLF}$$

and write $\mathcal{R}_{\mathsf{rLF}}(\epsilon, \mathcal{P}, \mathsf{B}.)$ for the region of (n, m)-values for which (rLF) can be performed successfully, defined analogously to $\mathcal{R}_{\mathsf{LF}}(\epsilon, \mathcal{P})$. Note that $\mathcal{R}_{\mathsf{rLF}} \subseteq \mathcal{R}_{\mathsf{LF}}$ provided $u \in \mathsf{B}_u$ for all $u \in \mathcal{P}$, that is, the range of sample sizes n, m for which robustly testing (LF) is possible ought to be a subset of $\mathcal{R}_{\mathsf{LF}}$.

Theorem 2.3.4. *Theorems 2.3.2 and 2.3.3 remain true if we replace $\mathcal{R}_{\mathsf{LF}}(\epsilon, \mathcal{P})$ by $\mathcal{R}_{\mathsf{rLF}}(\epsilon, \mathcal{P}, \mathsf{B}.)$ for the following choices:*

(i) for $\mathcal{P}_{\mathsf{H}}(\beta, d, C)$ and $\mathsf{B}_u = \{v \in \mathcal{P}_{\mathsf{H}}(\beta, d, C) : \|u - v\|_2 \leq c\epsilon\}$ for a constant $c > 0$ independent of ϵ,

(ii) for $\mathcal{P}_{\mathsf{G}}(s, C)$ and $\mathsf{B}_{\mu_\theta} = \{\mu_{\theta'} : \theta' \in \mathcal{E}(s, C), \|\theta - \theta'\|_2 \leq \epsilon/4\}$,

(iii) for $\mathcal{P}_{\mathsf{Db}}(k, C)$ and $\mathsf{B}_u = \{v : \|u - v\|_2 \leq \epsilon/(2\sqrt{k})\}$, and

(iv) for $\mathcal{P}_{\mathsf{D}}(k)$ and $\mathsf{B}_u = \{v : \|u - v\|_2 \leq c\epsilon/\sqrt{k}, \|v/u\|_\infty \leq c\}$ for a constant $c > 0$ independent of k and ϵ.

2.3.4 Beyond total variation

Recall from Remark 14 the notation $n_{\mathsf{GoF}}(\epsilon, \mathsf{d}, \mathcal{P})$ etc. that is applicable when separation is measured with respect to a general measure of discrepancy d instead of TV. In recent work [154, Theorem 1] and [164, Lemma 3.6] it is shown that any test that first quantizes the data by a map $\Phi : \mathcal{X} \to \{1, 2, \dots, M\}$ for some $M \geq 2$ must decrease the Hellinger distance between the two hypotheses by a log factor in the worst case. This implies that for every class $\mathcal{P}$ rich enough to contain such worst case examples, a quantizing test, such as Scheffé's, can hope to achieve $m \asymp \log(1/\epsilon)/\epsilon^2$ at best, as opposed to the optimal $m \asymp 1/\epsilon^2$. Thus, if separation is assumed with respect to Hellinger distance, Scheffé's test should be avoided. This example shows that the choice of d can have surprising effects on the performance of specific tests that would be optimal under other circumstances. Understanding the sample complexity of (LF) for d other than TV might lead to new algorithms and insights.

This motivates us to pose the question: does a trade-off analogous to that identified in Theorem 2.3.2 hold for other choices of d, and H in particular? In the case of $\mathcal{P}_{\mathsf{G}}$ we obtain a simple, almost vacuous answer. From Lemma 2.3.5 it follows immediately that the results of Table 2.1 and Theorem 2.3.2 continue to hold for $\mathcal{P}_{\mathsf{G}}$ for any of $\mathsf{d} \in \{\mathsf{H}, \sqrt{\mathsf{KL}}, \sqrt{\chi^2}\}$, to name a few.

Lemma 2.3.5. *Let $C > 0$ be a constant. For any $\theta \in \ell^2$ with $\|\theta\|_2 \leq C$*

$$\mathsf{TV}(\mu_\theta, \mu_0) \asymp \mathsf{H}(\mu_\theta, \mu_0) \asymp \sqrt{\mathsf{KL}(\mu_\theta \| \mu_0)} \asymp \sqrt{\chi^2(\mu_\theta \| \mu_0)} \asymp \|\theta\|_2,$$

where $\mu_\theta =: \otimes_{i=1}^{\infty} \mathcal{N}(\theta_i, 1)$ and the implied constant depends on C.

The case of $\mathcal{P}_{\mathsf{D}}$ is more intricate. Substantial recent progress [64, 122, 54, 35] has been made, where among others, the complexities $n_{\mathsf{GoF}}, n_{\mathsf{TS}}, n_{\mathsf{Est}}$ for Hellinger separation are identified. Since our algorithm for (LF) is $\|\cdot\|_2$-based, we could immediately derive achievability bounds

	n_{HT}	n_{GoF}	n_{TS}	n_{Est}
$\mathcal{P}_{\mathsf{D}}$	$1/\epsilon^2$	$\sqrt{k}/\epsilon^2$	$k^{2/3}/\epsilon^{8/3} \wedge k^{3/4}/\epsilon^2$	$n_{\mathsf{Gof}}^2\, \epsilon^2$
$\mathcal{P}_{\mathsf{H}}$	$1/\epsilon^2$	?	?	$1/\epsilon^{2(\beta+d)/\beta}$

Table 2.2: Prior results for $\mathsf{d} = \mathsf{H}$.

for $\mathcal{R}_{\mathsf{LF}}(\epsilon, \mathsf{H}, \mathcal{P}_{\mathsf{D}})$ via the inequality $\|\cdot\|_2 \geq \mathsf{H}^2/\sqrt{k}$, however such a naive technique yields suboptimal results, and thus we omit it. Studying (LF) under Hellinger separation for $\mathcal{P}_{\mathsf{D}}$ and $\mathcal{P}_{\mathsf{Db}}$ is beyond the scope of this work.

Finally, we turn to $\mathcal{P}_{\mathsf{H}}$. Due to the nature of our proofs, the results of Theorem 2.3.2 easily generalize to $\mathsf{d} = \|\cdot\|_p$ for any $p \in [1, 2]$. The simple reason for this is that (i) our algorithm is $\|\cdot\|_2$-based and $\|\cdot\|_2 \geq \|\cdot\|_p$ by Jensen's inequality and (ii) the lower bound construction involves perturbations near 1, where all said norms are equivalent. In the important case $\mathsf{d} = \mathsf{H}$ the estimation rate $n_{\mathsf{Est}}(\epsilon, \mathsf{H}, \mathcal{P}_{\mathsf{H}}) \asymp 1/\epsilon^{2(\beta+d)/\beta}$ was obtained by Birgé [26], our contribution here is the study of n_{GoF}.

Theorem 2.3.6. *For any $\beta > 0, C > 1$ and $d \geq 1$ there exists a constant $c > 0$ such that*

$$n_{\mathsf{GoF}}(\epsilon, \mathsf{H}, \mathcal{P}(\beta, d, C)) \geq c/\epsilon^{2(\beta+d/2)/\beta}.$$

If in addition we assume that $\beta \in (0, 1]$, c can be chosen such that

$$c n_{\mathsf{GoF}}(\epsilon, \mathsf{H}, \mathcal{P}) \leq 1/\epsilon^{2(\beta+d/2)/\beta}.$$

In particular, $n_{\mathsf{Est}} \asymp n_{\mathsf{GoF}}^2 \epsilon^2$.

2.4 Sketch proof of main results

In this section we briefly sketch the proofs of the main results of the paper.

2.4.1 Upper bounds for Theorems 2.3.2 to 2.3.4 and 2.3.6

Bounded discrete distributions

Consider first the case when $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ belong to the class $\mathcal{P}_{\mathsf{Db}}$, that is, they are supported on the discrete set $\{1, 2, \ldots, k\}$ and bounded by the uniform distribution. Let $\widehat{p}_{\mathsf{X}}, \widehat{p}_{\mathsf{Y}}, \widehat{p}_{\mathsf{Z}}$ denote empirical probability mass functions based on the samples X, Y, Z of size n, n, m from $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Z}}$ respectively. Define the test statistic

$$T_{\mathsf{LF}} = \|\widehat{p}_{\mathsf{X}} - \widehat{p}_{\mathsf{Z}}\|_2^2 - \|\widehat{p}_{\mathsf{Y}} - \widehat{p}_{\mathsf{Z}}\|_2^2$$

and the corresponding test $\psi(X, Y, Z) = \mathbb{1}\{T_{\mathsf{LF}} \geq 0\}$. The proof of Theorems 2.3.2 and 2.3.3 hinge on the precise calculation of the mean and variance of T_{LF}. Due to symmetry it is enough to compute these under the null. The proof of the upper bound is then completed via Chebyshev's inequality: if n, m are such that $(\mathbb{E}T_{\mathsf{LF}})^2 \gtrsim \mathrm{var}(T_{\mathsf{LF}})$ for large enough implied constant on the right then ψ tests (LF) successfully in the sense of (2.2.1).

Proposition 2.4.1 (informal)**.** *Suppose $\|p_{\mathsf{X}} + p_{\mathsf{Y}} + p_{\mathsf{Z}}\|_\infty \leq C_\infty/k$. Then ψ successfully tests* (LF) *if*

$$\underbrace{\frac{\epsilon^4}{k^2}}_{(\mathbb{E}T_{\mathsf{LF}})^2} \gtrsim \underbrace{\frac{C_\infty \epsilon^2}{k^2}\left(\frac{1}{n} + \frac{1}{m}\right) + \frac{C_\infty}{k}\left(\frac{1}{n^2} + \frac{1}{nm}\right)}_{\mathrm{var}(T_{\mathsf{LF}})}. \tag{2.4.1}$$

From (2.4.1) one can immediately see where each constraint in the region $\mathcal{R}_{\mathsf{LF}}(\epsilon, \mathcal{P}_{\mathsf{Db}}(k, C))$ in Theorem 2.3.2 emerges. The first two terms in the variance require that both m and n be larger than $\Omega(1/\epsilon^2)$. The $1/n^2$ term in the variance requires that n be at least $\Omega(\sqrt{k}/\epsilon^2) \asymp n_{\mathsf{GoF}}$, and the $1/(nm)$ term requires that the product nm be at least $\Omega(n_{\mathsf{GoF}}^2)$.

Smooth densities

Next we describe how Proposition 2.4.1 can be applied to the class $\mathcal{P}_{\mathsf{H}}$ of smooth densities. Divide $[0, 1]^d$ into into κ^d regular grid cells for some $\kappa \in \mathbb{N}$. Discretize the three samples X, Y, Z over this grid and simply apply the optimal test for $\mathcal{P}_{\mathsf{Db}}$, observing the crucial fact that this discretization belongs to $\mathcal{P}_{\mathsf{Db}}$. The following lemma, originally due to Ingster [110] controls the approximation error of the discretization.

Lemma 2.4.2 ([10, Lemma 7.2]). *Let P_κ denote the L^2 projection onto the space of functions constant on each grid cell. For any $\beta > 0, C > 1$ and $d \geq 1$ there exist constants $c, c' > 0$ such that for any $f, g \in \mathcal{P}_{\mathsf{H}}(\beta, d, C)$ the following holds:*

$$\|f - g\|_2 \geq \|P_\kappa(f - g)\|_2 \geq c\|f - g\|_2 - c'\kappa^{-\beta}.$$

Based on Lemma 2.4.2 we set $\kappa \asymp \epsilon^{-1/\beta}$. This resolution is chosen to ensure that the discrete approximation to any β-smooth density is sufficiently accurate, that is, approximate ϵ-separation is maintained even after discretization. We see now that our problem is reduced entirely to testing over $\mathcal{P}_{\mathsf{Db}}$, so we may apply Proposition 2.4.1 with $k = \kappa^d \asymp \epsilon^{-d/\beta}$, which yields the minimax optimal rates from Theorems 2.3.2 and 2.3.4.

Our proof of the achieavability direction in Theorem 2.3.6 follows similarly by reduction to goodness-of-fit testing for discrete distributions [54] under Hellinger separation, where it is known that $n_{\mathsf{GoF}}(\epsilon, \mathsf{H}, \mathcal{P}_{\mathsf{D}}) \asymp \sqrt{k}/\epsilon^2$. The key step is to prove a result similar to Lemma 2.4.2 but for H instead of $\|\cdot\|_2$.

Proposition 2.4.3. *For any $\beta \in (0, 1]$, $C > 1$ and $d \geq 1$ there exists a constant $c > 0$ such that*

$$c\mathsf{H}(f, g) \leq \mathsf{H}(P_\kappa f, P_\kappa g) \leq \mathsf{H}(f, g)$$

holds for any $f, g \in \mathcal{P}_{\mathsf{H}}(\beta, d, C)$, provided we set $\kappa = (c\epsilon)^{-2/\beta}$.

Gaussian sequence model

Let us briefly discuss the Gaussian sequence class $\mathcal{P}_{\mathsf{G}}(s, C)$. Here our approach is not to discretize the distributions, but conceptually the test is very similar to the cases we've already covered. Let us write $\mathbb{P}_{\mathsf{X}} = \mu_{\theta_{\mathsf{X}}}$ and define $\theta_{\mathsf{Y}}, \theta_{\mathsf{Z}}$ analogously. For a given cutoff r, we simply

$$\text{reject the null if } T_{\mathsf{LF},\mathsf{G}} =: \sum_{i=1}^{r} \left\{ \left(\widehat{\theta}_{\mathsf{X},i} - \widehat{\theta}_{\mathsf{Z},i}\right)^2 - \left(\widehat{\theta}_{\mathsf{Y},i} - \widehat{\theta}_{\mathsf{Z},i}\right)^2 \right\} \geq 0, \tag{2.4.2}$$

where $\widehat{\theta}_{\mathsf{X},i} = \frac{1}{n}\sum_{j=1}^{n} X_{ji}$ and $\widehat{\theta}_{\mathsf{Y}}, \widehat{\theta}_{\mathsf{Z}}$ are defined analogously. Once again, a precise calculation of the mean and the variance of the sum above, yields the following result.

Proposition 2.4.4 (informal). *Set $r \asymp \epsilon^{-1/s}$. The test (2.4.2) succeeds if*

$$\underbrace{\epsilon^4}_{(\mathbb{E}T_{\mathsf{LF},\mathsf{G}})^2} \gtrsim \underbrace{\epsilon^2\left(\frac{1}{n} + \frac{1}{m}\right) + \epsilon^{-1/s}\left(\frac{1}{n^2} + \frac{1}{nm}\right)}_{\mathrm{var}(T_{\mathsf{LF},\mathsf{G}})}. \tag{2.4.3}$$

Similarly to (2.4.1), we can again read of the constraints that define the region $\mathcal{R}_{\mathsf{LF}}(\epsilon, \mathcal{P}_{\mathsf{G}}(s, C))$ from (2.4.3). The first and second terms in the variance ensure that $n, m = \Omega(1/\epsilon^2)$ and $n^2, mn = \Omega(n_{\mathsf{GoF}}^2) = \Omega(\epsilon^{-(4s+1)/s})$ respectively.

General discrete distributions

Finally, we comment on $\mathcal{P}_{\mathsf{D}}$. Here we can no longer assume that $C_\infty = \mathcal{O}(1)$ in Proposition 2.4.1, in fact $C_\infty = \Omega(k)$ is possible. We get around this by utilizing the reduction based approach of [64, 80]. We take the first half of the data and compute

$$B_i = 1 + \#\left\{j \le \frac{\min\{k,n\}}{2} : X_j = i\right\} + \#\left\{j \le \frac{\min\{k,n\}}{2} : Y_j = i\right\} + \#\left\{j \le \frac{\min\{k,m\}}{2} : Z_j = i\right\}$$

for each $i \in [k]$. Then, we divide the i'th support element into B_i bins, uniformly. This transformation preserves pairwise total variation, but reduces the ℓ^∞-norms of $p_{\mathsf{X}}, p_{\mathsf{Y}}, p_{\mathsf{Z}}$ with high probability, to order $1/(k \wedge (n \vee m))$, after an additional step that we omit here. We can then perform the usual test with these new "flattened" distributions, using the untouched half of the data.

It is insightful to interpret the "flattening" procedure followed by L^2-distance comparison as a one-step procedure that simply compares a different divergence of the empirical measures. Intuitively, in contrast to the regular classes, one needs to mitigate the effect of potentially massive differences in the empirical counts on bins $i \in [k]$ where both $p_{\mathsf{X}}(i)$ and $p_{\mathsf{Y}}(i)$ are large but their difference $|p_{\mathsf{X}}(i) - p_{\mathsf{Y}}(i)|$ is moderate. Let LC_λ be the "weighted Le-Cam divergence" which we define as $\mathsf{LC}_\lambda(p\|q) = \sum_i (p_i - q_i)^2/(p_i + \lambda q_i)$ for two probability mass functions p, q. One may interpret the two step procedure (flattening followed by comparing L^2 distances) as approximately comparing empirical weighted Le-Cam divergences. Performing the test in two steps is a proof device, and we expect the test that directly compares, say, the Le-Cam divergence of the empirical probability mass functions to have the same minimax optimal sample complexity. Such a one-shot approach is used for example in the paper [39] for two-sample testing. While Ingster [110] only considers goodness-of-fit testing to the uniform distribution, his notation also suggests the idea of normalizing by the bin mass under the null.

2.4.2 Lower bounds for Theorems 2.3.2 to 2.3.4 and 2.3.6

The reductions given in Proposition 2.3.1 immediately yields a number of tight lower bounds on n and m. Namely, (2.3.1) gives $m \gtrsim 1/\epsilon^2$ and (2.3.3) gives $n \gtrsim n_{\mathsf{GoF}}$. Obtaining the lower bound on the product term mn proves more challenging. First we introduce the well known information theoretic tools we use to prove our minimax lower bounds.

Suppose that we have two (potentially composite) hypotheses H_0, H_1 that we test against each other. Our strategy relies on the method of two fuzzy hypotheses [198], which is a generalization of Le-Cam's two point method. Write $\mathcal{M}(\mathcal{X})$ for the set of probability measures on the set $\mathcal{X}$.

Lemma 2.4.5. *Take two hypotheses $H_i \subseteq \mathcal{M}(\mathcal{X})$ and random $P_i \in \mathcal{M}(\mathcal{X})$. Then*

$$2 \inf_\psi \max_{i=0,1} \sup_{P \in H_i} P(\psi \ne i) \ge 1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) - \sum_i \mathbb{P}(P_i \notin H_i),$$

where the infimum is over all tests $\psi : \mathcal{X} \to \{0,1\}$.

Proof. We may assume without loss of generality that $\mathbb{P}(P_i \in H_i) > 0$ for both $i = 0$ and $i = 1$, as otherwise the claim is vacuous. Let $\tilde{P}_i$ be distributed as $P_i|\{P_i \in H_i\}$. Then for any set $A \subset \mathcal{X}$ we have

$$\left|\mathbb{E}\tilde{P}_i(A) - \mathbb{E}P_i(A)\right| = \mathbb{P}(P_i \notin H_i)\left|\mathbb{E}[P_i(A)|P_i \in H_i] - \mathbb{E}[P_i(A)|P_i \notin H_i]\right| \le \mathbb{P}(P_i \notin H_i).$$

In particular, $\mathsf{TV}(\mathbb{E}\tilde{P}_0, \mathbb{E}\tilde{P}_1) \le \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) + \sum_i \mathbb{P}(P_i \notin H_i)$. Therefore, for any ψ

$$\max_{i=0,1} \sup_{\mathbb{P}_i \in H_i} \mathbb{P}_i(\psi \neq i) \ge \frac{1}{2}(1 - \mathsf{TV}(\mathbb{E}\tilde{P}_0, \mathbb{E}\tilde{P}_1)) \ge \frac{1}{2}\left(1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) - \sum_i \mathbb{P}(P_i \notin H_i)\right).$$

□

For clarity, we formally state (LF) as testing between the hypotheses

$$\begin{aligned} H_0 &= \{\mathbb{P}_{\mathsf{X}}^{\otimes n} \otimes \mathbb{P}_{\mathsf{Y}}^{\otimes n} \otimes \mathbb{P}_{\mathsf{X}}^{\otimes m} : \mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P},\ \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \ge \epsilon\} \\ &\quad \text{versus} \\ H_1 &= \{\mathbb{P}_{\mathsf{X}}^{\otimes n} \otimes \mathbb{P}_{\mathsf{Y}}^{\otimes n} \otimes P_{\mathsf{Y}}^{\otimes m} : \mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P},\ \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \ge \epsilon\}. \end{aligned} \tag{2.4.4}$$

The lower bounds of Theorem 2.3.4 follow from those for Theorems 2.3.2 and 2.3.3 so we may focus on the latter.

Smooth densities

For concreteness let us focus on the case of $\mathcal{P} = \mathcal{P}_{\mathsf{H}}$. We take $\mathbb{P}_0$ to be uniform on $[0,1]^d$ and $\mathbb{P}_\eta$ to have density

$$p_\eta = 1 + \sum_{j \in [\kappa]^d} \eta_j h_j \tag{2.4.5}$$

with respect to $\mathbb{P}_0$. Here $\kappa \in \mathbb{N}$, each $\eta \in \{\pm 1\}^{\kappa^d}$ is uniform and h_j is a bump function supported on the j'th cell of the regular grid of size κ^d on $[0,1]^d$. The parameters κ, h_j of the construction are set in a way to ensure $\mathbb{P}_\eta \in \mathcal{P}_{\mathsf{H}}$ and $\mathsf{TV}(\mathbb{P}_0, \mathbb{P}_\eta) \ge \epsilon$ with probability 1 over η. We have

$$\begin{aligned} 1 + \chi^2(\mathbb{E}_\eta \mathbb{P}_\eta^{\otimes m} \| \mathbb{P}_0^{\otimes m}) &= \int_{[0,1]^{dm}} \left(\mathbb{E}_\eta \prod_{i=1}^n p_\eta(x_i)\right)^2 \mathrm{d}x_1 \dots \mathrm{d}x_m \\ &= \mathbb{E}_{\eta\eta'} \langle p_\eta, p_{\eta'} \rangle_{L^2}^m \\ &= \mathbb{E}(1 + \|h_{\mathbb{1}}\|_2^2 \langle \eta, \eta' \rangle)^m \\ &\le \exp(m^2 \|h_{\mathbb{1}}\|_2^4 \kappa^d), \end{aligned} \tag{2.4.6}$$

where η, η' are i.i.d. uniform and we assume $\|h_{\mathbb{1}}\|_2 = \|h_j\|_2$ for all $j \in [\kappa]^d$. The above approach is what Ingster used in his seminal paper [110] on goodness-of-fit testing, which we adapt to likelihood-free hypothesis testing (2.4.4). Take $P_0 = \mathbb{P}_\eta^{\otimes n} \otimes \mathbb{P}_0^{\otimes n} \otimes \mathbb{P}_\eta^{\otimes m}$ and $P_1 = \mathbb{P}_\eta^{\otimes n} \otimes \mathbb{P}_0^{\otimes n} \otimes \mathbb{P}_0^{\otimes m}$ in Lemma 2.4.5. Bounding $\mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1)$ proceeds in multiple steps: first, we drop the Y-sample using the data-processing inequality. Then, we use Pinsker's

inequality and the chain rule to bound TV by the KL divergence of Z conditioned on X. We bound KL by χ^2, arriving at the same equation (2.4.6). However, the mixing parameters η, η' are no longer independent, instead, given X they're independent from the posterior. In the remaining steps we use the fact that the posterior factorizes over the bins and the calculation is reduced to just a single bin where it can be done explicitly.

Let us now turn to the lower bound in Theorem 2.3.6. The difference in the rate is a consequence of the fact that H and TV behave differently for densities near zero. Inspired by this, we slightly modify the construction (2.4.5) by putting the perturbations at density level ϵ^2 as opposed to 1. Bounding TV then proceeds analogously to the steps outlined above.

Bounded discrete distributions

The construction is entirely analogous to the case of $\mathcal{P}_{\mathsf{H}}$ and we refer to the appendix for details. In the computer science community the construction of p_η is attributed to Paninski [162].

Gaussian sequence model

The null distribution $\mathbb{P}_0$ is the no signal case $\otimes_{i=1}^{\infty}\mathcal{N}(0,1)$ while the alternative is $\mathbb{P}_\theta = \otimes_{i=1}^{\infty}\mathcal{N}(\theta_i, 1)$ where θ has prior distribution $\otimes_{i=1}^{\infty}\mathcal{N}(0, \gamma_i)$ for an appropriate sequence $\gamma \in \mathbb{R}^{\mathbb{N}}$. We refer to the appendix for more details.

General discrete distributions

Once again, the irregular case $\mathcal{P}_{\mathsf{D}}$ requires special consideration. Clearly the lower bound for $\mathcal{P}_{\mathsf{Db}}$ carries over. However, in the regime $k \gtrsim 1/\epsilon^4$ said lower bound becomes suboptimal, and we need a new construction, for which we utilize the moment-matching based approach of Valiant [200] as a black-box. The construction is derived from that used for two-sample testing by Valiant, namely the pair $(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}})$ is chosen uniformly at random from $\{(p \circ \pi, q \circ \pi)\}_{\pi \in S_k}$. Here we write S_k for the symmetric group on $[k]$ and

$$p(i) = \begin{cases} \frac{1-\epsilon}{n} & \text{for } i \in [n] \\ \frac{4\epsilon}{k} & \text{for } i \in [\frac{k}{2}, \frac{3k}{4}] \\ 0 & \text{otherwise,} \end{cases}$$

where we assume that $m \le n \le k/2$ and define $q(i) = p(i)$ for $i \in [k/2 - 1]$ and $q(i) = p(3k/2 - i)$ for $i \in [k/2, k]$. This construction gives a lower bound matching our upper bound in the regime $m \lesssim n \lesssim k$. The final piece of the puzzle follows by the reduction from two-sample testing with unequal sample size (2.3.6), as this shows that likelihood-free hypothesis testing is at least as hard as two-sample testing in the $n \le m$ regime, and known lower bounds on the sample complexity of two-sample testing [25] (see also Table 2.1) let us conclude.

2.5 Open problems

A natural follow-up direction to the present paper would be to study multiple hypothesis testing where $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ are replaced by $\mathbb{P}_{\mathsf{X}_1}, \dots, \mathbb{P}_{\mathsf{X}_M}$ with corresponding hypotheses $H_1, \dots, H_M$. The geometry of the family $\{\mathbb{P}_{\mathsf{X}_j}\}_{j\in[M]}$ might have interesting effects on the sample complexities.

Open problem 1. *Study the dependence on $M > 2$ of likelihood-free testing with M hypotheses.*

Another possible avenue of research is the study of local minimax/instance optimal rates, which is the focus of recent work [199, 13, 45, 44, 137] in the case of goodness-of-fit and two-sample testing.

Open problem 2. *Define and study the local minimax rates of likelihood-free hypothesis testing.*

Our discussion of the Hellinger case in Section 2.3.4 is quite limited, natural open problems in this direction include the following.

Open problem 3. *Let $\mathcal{P} \in \{\mathcal{P}_{\mathsf{H}}(\beta, d, C), \mathcal{P}_{\mathsf{Db}}(k, C_{\mathsf{Db}}), \mathcal{P}_{\mathsf{D}}(k)\}$.*

(i) Study n_{GoF} and n_{TS} for $\mathcal{P}$ under Hellinger separation.

(ii) Determine the trade-off $\mathcal{R}_{\mathsf{LF}}$ for $\mathcal{P}$ under Hellinger separation.

More ambitiously, one might ask for a characterization of 'regular' models $(\mathcal{P}, \mathsf{d})$ for which goodness-of-fit testing and two-sample testing are equally hard and the region $\mathcal{R}_{\mathsf{LF}}$ is given by the trade-off in Theorem 2.3.2.

Open problem 4. *Find a general family of 'regular' models $(\mathcal{P}, \mathsf{d})$ for which*

$$n_{\mathsf{GoF}}(\epsilon, \mathsf{d}, \mathcal{P}) \asymp n_{\mathsf{TS}}(\epsilon, \mathsf{d}, \mathcal{P}) \text{ and}$$
$$\mathcal{R}_{\mathsf{LF}}(\epsilon, \mathsf{d}, \mathcal{P}) \asymp \{m \geq 1/\epsilon^2, n \geq n_{\mathsf{GoF}}(\epsilon, \mathsf{d}, \mathcal{P}), mn \geq n^2_{\mathsf{GoF}}(\epsilon, \mathsf{d}, \mathcal{P})\}.$$

Recent follow-up work [74] showed that Scheffé's test is also minimax optimal and achieves the entire trade-off in Figure 2.1. It appears that the optimality of Scheffé's test is a consequence of the minimax point of view. Basically, in the worst-case the log-likelihood ratio between the hypotheses is close to being binary, hence quantizing it to $\{0, 1\}$ does not lose optimality. Consequently, an important future direction is to better understand the competitive properties of various tests and studying some notion of regret, see [2] for prior related work.

Open problem 5. *Study the competitive optimality of likelihood-free hypothesis testing algorithms, and Scheffé's test in particular.*

Chapter 3

Kernel-Based Tests for Likelihood-Free Hypothesis Testing

This chapter is a reproduction of [76], which was published at the Thirty Seventh Annual Conference on Neural Information Processing Systems, and is joint work with Tianze Jiang, Yury Polyanskiy and Rui Sun.

3.1 Likelihood-Free Inference

The goal of likelihood-free inference (LFI) [66, 91, 32, 51], also called simulation-based inference (SBI), is to perform statistical inference in a setting where the data generating process is a black-box, but can be simulated. Given the ability to generate samples $X_\theta \sim P_\theta^{\otimes n}$ for any parameter θ, and given real-world data $Z \sim P_{\theta^\star}^{\otimes m}$, we want to use our simulations to learn about the truth $\theta^\star$. LFI is particularly relevant in areas of science where we have precise but complex laws of nature, for which we can do (stochastic) forward simulations, but can not directly compute the (distribution) density P_θ. The Bayesian community approached the problem under the name of Approximate Bayesian Computation (ABC) [52, 182, 22]. More recent ML-based methods where regressors and classifiers are used to summarize data, select regions of interest, approximate likelihoods or likelihood-ratios [114, 117, 163, 53, 196] have also emerged for this challenge.

Despite empirical advances, the theoretical study of frequentist LFI is still in its infancy. We focus on the nonparametric and non-asymptotic setting, which we justify as follows. For applications where tight error control is critical one might be reluctant to rely on asymptotics. More broadly, the non-asymptotic regime can uncover new phenomena and provide insights for algorithm design. Further, parametric models are clearly at odds with the black-box assumption. Recently, [78] proposed likelihood-free hypothesis testing (LFHT) as a simplified model and found minimax optimal tests for a range of nonparametric distribution classes, thereby identifying a *fundamental simulation-experimentation trade-off* between the number of simulated observations n and the size of the experimental data sample m. Here we extend [78], and prior related work [102, 100, 124, 123, 74], to a new setting designed to model experimental setups more truthfully and derive sample complexity (upper and lower bounds) for kernel-based tests over nonparametric classes.

While minimax optimal, the algorithms of [78, 74] are impractical as they rely on discretizing the observations on a regular grid. Thus, both in our theory as well as experiments we turn to kernel methods which provide an empirically more powerful set of algorithms that have shown success in nonparametric testing [86, 85, 118, 87, 143].

Contributions Our contributions are twofold. *Theoretically*, we introduce *mixed likelihood-free hypothesis testing* (mLFHT), which is a generalization of (LFHT) and provides a better model of applications such as the search for new physics [49, 142]. We propose a robust kernel-based test and derive both upper and lower bounds on its minimax sample complexity over a large nonparametric class of densities, generalizing multiple results in [123, 102, 100, 124, 139, 78, 74]. Although the simulation-experimentation (m vs n) trade-off has been proven in the minimax sense (that is, for some worst-case data distribution), it is not clear whether it actually occurs in real data. Our second contribution is the *empirical* confirmation of the existence of an asymmetric trade-off, cf. Figure 3.1. To this end we construct state-of-the-art tests building on ideas of [187, 143] on learning good kernels from the data. We execute this program in two settings: the Higgs boson discovery [14], and detecting diffusion [95] generated images planted in the CIFAR-10 [135] dataset.

3.1.1 LFHT and the Simulation-Experimentation Trade-off

Suppose we have i.i.d. samples X, Y each of size n from two unknown distributions P_X, P_Y on a measurable space $\mathcal{X}$, as well as a third i.i.d. sample $Z \sim P_Z$ of size m. In the context of LFI, we may think of the samples X, Y as being generated by our simulator, and Z being the data collected in the real world. The problem we refer to as likelihood-free hypothesis testing is the task of deciding between the hypotheses

$$H_0 : P_Z = P_X \quad \text{versus} \quad H_1 : P_Z = P_Y. \tag{LFHT}$$

This problem originates in [90, 212], where authors study the exponents of error decay for finite $\mathcal{X}$ and fixed P_X, P_Y as $n \sim m \to \infty$; more recently [123, 102, 100, 124, 2, 139, 78, 74] it is studied in the non-asymptotic regime. Assuming that P_X, P_Y belong to a known nonparametric class of distributions $\mathcal{P}$ and are guaranteed to be ϵ-separated with respect to total variation (TV) distance (i.e. $\mathsf{TV}(P_X, P_Y) \geq \epsilon$), [78] characterizes the sample sizes n and m required for the sum of type-I and type-II errors to be small, as a function of ϵ and for several different $\mathcal{P}$'s. Their results show, for three settings of $\mathcal{P}$, that (i) testing (LFHT) at vanishing error is possible even when n is not large enough to estimate P_X and P_Y within total variation distance $\mathcal{O}(\epsilon)$, and that (ii) to achieve a fixed level of error, say α, one can *trade off* m vs. n along a curve of the form $\{\min\{n, \sqrt{mn}\} \gtrsim n_{\mathsf{TS}}(\alpha, \epsilon, \mathcal{P}), m \gtrsim \log(1/\alpha)/\epsilon^2\}$. Here n_{TS} denotes the minimax sample complexity of two-sample testing over $\mathcal{P}$, i.e. the minimum number of observations n needed from $P_X, P_Y \in \mathcal{P}$ to distinguish the cases $\mathsf{TV}(P_X, P_Y) \geq \epsilon$ versus $P_X = P_Y$. Here $\gtrsim$ suppresses dependence on constants and untracked parameters.

It is unclear, however, whether predictions drawn from minimax sample complexities over specified distribution classes can be observed in real-world data. Without the theory, a natural expectation is that the error contour $\{(m, n) : \exists$ a test with total error $\leq \alpha\}$ would look similar to that of minimax two-sample testing with unequal sample size, namely $\{(m, n) : \min\{n, m\} \gtrsim n_{\mathsf{TS}}(\alpha, \epsilon, \mathcal{P})\}$, i.e. n and m simply need to be above a certain threshold

simultaneously (as is the case for e.g. two-sample testing over smooth densities [10, 139]). However, from Figures 3.2 and 3.1 we see that there is indeed a non-trivial trade-off between n and m: the contours are not always parallel to the axes and aren't symmetric about the line $m = n$. The importance of Fig. 3.1 is in demonstrating that said trade-off is not a kink of a theory that arises due to some esoteric worst-case data distribution, but is instead a real effect observed in state-of-the-art LFI algorithms ran on actual data. We remark that the n used in this plot is the total number of simulated samples (most of which are used for choosing a neural-network parameterized kernel) and are not just the n occuring in Theorems 3.3.1 and 3.3.2 which apply to a *fixed* kernel. See Section 3.4 for details on sample division.

3.1.2 Mixed Likelihood-Free Hypothesis Testing

A prominent application of likelihood-free inference lies in the field of particle physics. Scientists run sophisticated experiments in the hope of finding a new particle or phenomenon. Often said phenomenon can be predicted from theory, and thus can be simulated, as was the case for the Higgs boson whose existence was verified after nearly 50 years at the Large Hadron Collider (LHC) [40, 5].

Suppose we have n simulations from the *background* distribution P_X and the *signal* distribution P_Y. Further, we also have m (real-world) datapoints from $P_Z = (1-\nu)P_X + \nu P_Y$, i.e. the observed data is a mixture between the background and signal distributions with rate parameter ν. The goal of physicists is to construct confidence intervals for ν, and a *discovery* corresponds to a 5σ confidence interval that excludes $\nu = 0$. We model this problem by testing

$$H_0 : \nu = 0 \quad \text{versus} \quad H_1 : \nu \geq \delta \tag{mLFHT}$$

for fixed (usually predicted) $\delta > 0$. See the rigorous definition of (mLFHT) in Section 3.3. In particular, a discovery can be claimed if H_0 is rejected.

3.2 The Likelihood-Free Test Statistic

This section introduces the testing procedure based on Maximum Mean Discrepancy (MMD) that we study throughout the paper both theoretically and empirically. First, we introduce the necessary background on MMD in Section 3.2.1. Then, we define our test statistics in Section 3.2.2.

3.2.1 Kernel Embeddings and MMD

Given a set $\mathcal{X}$, we call the function $K : \mathcal{X}^2 \to \mathbb{R}$ a kernel if the $n \times n$ matrix with ij'th entry $K(x_i, x_j)$ is symmetric positive semidefinite for all choices of $x_1, \ldots, x_n \in \mathcal{X}$ and $n \geq 1$. There is a unique reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ associated to K. $\mathcal{H}_K$ consists of functions $\mathcal{X} \mapsto \mathbb{R}$ and satisfies the reproducing property $\langle K(x, \cdot), f\rangle_{\mathcal{H}_K} = f(x)$ for all $f \in \mathcal{H}_k$ and $x \in \mathcal{X}$, in particular $K(x, \cdot) \in \mathcal{H}_K$. Given a probability measure P on $\mathcal{X}$, define its kernel embedding θ_P as

$$\theta_P \coloneqq \mathbb{E}_{X\sim P} K(X, \cdot) = \int_{\mathcal{X}} K(x, \cdot) P(\mathrm{d}x). \tag{3.2.1}$$

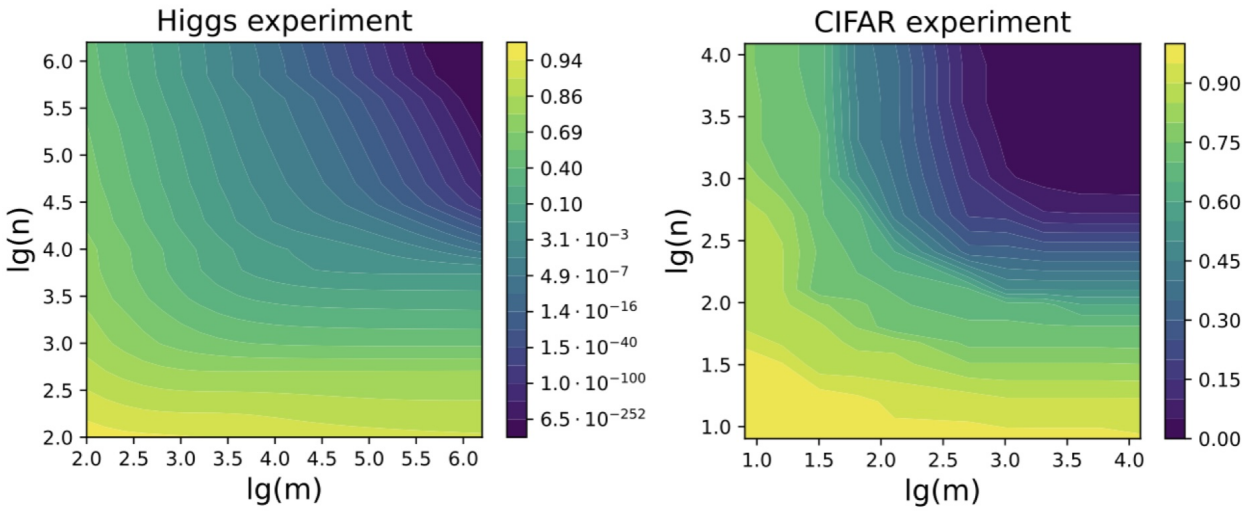


Figure 3.1: n versus m trade-off for the Higgs and CIFAR experiments using our test in Section 3.2. Error probabilities are estimated by normal approximation for Higgs and simulated for CIFAR.

Given the kernel embeddings of two probability measures P, Q, we can measure their distance in the RKHS by $\mathrm{MMD}(P,Q) := \|\theta_P - \theta_Q\|_{\mathcal{H}_K}$, where MMD stands for maximum mean discrepancy. MMD has a closed form thanks to the reproducing property and linearity:

$$\mathrm{MMD}^2(P,Q) = \mathbb{E}\Big[K(X,X') + K(Y,Y') - 2K(X,Y)\Big]$$

where $(X, X', Y, Y') \sim P^{\otimes 2} \otimes Q^{\otimes 2}$. In particular, if P, Q are empirical measures based on observations, we can evaluate the MMD exactly, which is crucial in practice. Yet another attractive property of MMD is that (under mild integrability conditions) it is an integral probability metric (IPM) where the supremum is over the unit ball of the RKHS $\mathcal{H}_K$. See e.g. [179, 151] for references. The following result is a consequence of the fact that self-adjoint compact operators are diagonalizable.

Theorem 3.2.1 (Hilbert–Schmidt)**.** *Suppose that $K \in L^2(\mu \otimes \mu)$ is symmetric. Then there exists a sequence $(\lambda_j)_{j\geq 1} \in \ell^2$ and an orthonormal basis $\{e_j\}_{j\geq 1}$ of $L^2(\mu)$ such that $K(x,y) = \sum_{j\geq 1} \lambda_j e_j(x) e_j(y)$ for all $j \geq 1$, where convergence is in $L^2(\mu \otimes \mu)$.*

Assumption 2. *Unless specified otherwise, we implicitly assume a choice of a non-negative measure μ and kernel K for which the conditions of Theorem 3.2.1 hold. Note that $\lambda_j \geq 0$ and depend on μ.*

Removing the Bias In our proofs we work with the kernel embedding of empirical measures for which we need to modify the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$ (and thus MMD) slightly by removing the diagonal terms. Namely, given i.i.d. samples X, Y of size n, m respectively and corresponding empirical measures $\widehat{P}_X, \widehat{P}_Y$, we define

$$\mathrm{MMD}_u^2(\widehat{P}_X, \widehat{P}_Y) := \sum_{i\neq j} \frac{K(X_i, X_j)}{n(n-1)} + \sum_{i\neq j} \frac{K(Y_i, Y_j)}{m(m-1)} - 2\sum_{i,j} \frac{K(X_i, Y_j)}{mn}. \tag{3.2.2}$$

We also write $\langle \theta_{\widehat{P}_X}, \theta_{\widehat{P}_X} \rangle_{u,\mathcal{H}_K} := \|\theta_{\widehat{P}_X}\|^2_{u,\mathcal{H}_K} := \frac{1}{n(n-1)} \sum_{i \neq j} K(X_i, X_j)$ and extend linearly. The u stands for unbiased, since $\mathbb{E}\,\mathrm{MMD}^2_u(\widehat{P}_X, \widehat{P}_Y) = \mathrm{MMD}^2(P_X, P_Y) \neq \mathbb{E}\,\mathrm{MMD}^2(\widehat{P}_X, \widehat{P}_Y)$ in general.

3.2.2 Test Statistic

With Section 3.2.1 behind us, we are in a position to define the test statistic that we use to tackle (mLFHT). Suppose that we have samples X, Y, Z of sizes n, n, m from the probability measures P_X, P_Y, P_Z. Write $\widehat{P}_X$ for the empirical measure of sample X, and analogously for Y, Z. The core of our test statistic for (mLFHT) is the following:

$$T(X,Y,Z) := \langle \theta_{\widehat{P}_Z}, \theta_{\widehat{P}_Y} - \theta_{\widehat{P}_X} \rangle_{u,\mathcal{H}_K} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \Big\{ K(Z_j, Y_i) - K(Z_j, X_i) \Big\}. \tag{3.2.3}$$

Note that T is of the additive form $\frac{1}{m}\sum_{j=1}^m f(Z_i)$ where $f(z) := \theta_{\widehat{P}_Y}(z) - \theta_{\widehat{P}_X}(z)$ can be interpreted as the *witness function* of [85, 118]. Given some $\pi \in [0,1]$ (taken to be half the predicted signal rate $\delta/2$ in our proofs), the output of our test is

$$\Psi_\pi = \mathbb{1}\Big\{ T(X,Y,Z) \geq \gamma(X,Y,\pi) \Big\}, \text{ where } \gamma(X,Y,\pi) = \pi\,\mathrm{MMD}^2_u(\widehat{P}_X, \widehat{P}_Y) + T(X,Y,X). \tag{3.2.4}$$

The threshold γ gives Ψ_π a natural geometric interpretation: it checks whether the projection of $\theta_{\widehat{P}_Z} - \theta_{\widehat{P}_X}$ onto the vector $\theta_{\widehat{P}_Y} - \theta_{\widehat{P}_X}$ falls further than π along the segment joining $\theta_{\widehat{P}_X}$ to $\theta_{\widehat{P}_Y}$ (up to deviations due to the omitted diagonal terms, see Section 3.2.1).

Setting $\delta = 1$ in (mLFHT) recovers (LFHT), and the corresponding test output is $\Psi_{\delta/2} = \Psi_{1/2} = 1$ if and only if $\mathrm{MMD}_u(\widehat{P}_Z, \widehat{P}_X) \geq \mathrm{MMD}_u(\widehat{P}_Z, \widehat{P}_Y)$. This very statistic (i.e. $\mathrm{MMD}_u(\widehat{P}_Z, \widehat{P}_X) - \mathrm{MMD}_u(\widehat{P}_X, \widehat{P}_Y)$) has been considered in the past for relative goodness-of-fit testing [31] where it's asymptotic properties are established. In the non-asymptotic setting, if MMD is replaced by the L^2-distance we recover the test statistic studied by [102, 124, 78]. However, we are the first to introduce Ψ_δ for $\delta \neq 1$ and to study MMD-based tests for (m)LFHT in a non-asymptotic setting.

Variance cancellation At first sight it may seem more natural to the reader to threshold the distance $\mathrm{MMD}_u(\widehat{P}_Z, \widehat{P}_X)$, resulting in rejection if, say, $\mathrm{MMD}_u(\widehat{P}_Z, \widehat{P}_X) \geq \mathrm{MMD}_u(\widehat{P}_X, \widehat{P}_Y)\delta/2$. The geometric meaning of this would be similar to the one outlined above. However, there is a crucial difference: (LFHT) (the case $\delta = 1$) is possible with very little experimental data m due to the *cancellation of variance*. More precisely, the statistic $\mathrm{MMD}^2(\widehat{P}_Z, \widehat{P}_X)$ contains the term $\frac{1}{m(m-1)}\sum_{i\neq j} K(Z_i, Z_j)$ — whose variance is prohibitively large and would inflate the m required for reliable testing — but this can be canceled by subtracting $\mathrm{MMD}^2(\widehat{P}_Z, \widehat{P}_Y)$. Our statistic $T(X,Y,Z) - \gamma(X,Y,\pi)$ simply generalizes this idea to (mLFHT).

3.3 Minimax Rates of Testing

3.3.1 Upper Bounds on the Minimax Sample Complexity of (mLFHT)

Let us start by reintroducing (mLFHT) in a rigorous fashion. Given $C, \epsilon, R \geq 0$, let $\mathcal{P}_\mu(C, \epsilon, R)$ denote the set of triples (P_X, P_Y, P_Z) of distributions such that the following three conditions hold:

(*i*) P_X, P_Y and P_Z have μ-densities bounded by C,

(*ii*) $\mathrm{MMD}(P_X, P_Y) \geq \epsilon$, and

(*iii*) $\mathrm{MMD}(P_Z, (1-\nu)P_X + \nu P_Y) \leq R \cdot \mathrm{MMD}(P_X, P_Y)$,

where we define $\nu = \nu(P_X, P_Y, P_Z) = \arg\min_{\nu' \in \mathbb{R}} \mathrm{MMD}(P_Z, (1-\nu')P_X + \nu' P_Y)$. For some $\delta > 0$, consider the two hypotheses

$$\begin{aligned} H_0(C, \epsilon, \delta, R) &: (P_X, P_Y, P_Z) \in \mathcal{P}_\mu(C, \epsilon, R) \text{ and } \nu = 0 \\ H_1(C, \epsilon, \delta, R) &: (P_X, P_Y, P_Z) \in \mathcal{P}_\mu(C, \epsilon, R) \text{ and } \nu \geq \delta, \end{aligned} \tag{3.3.1}$$

which we regard as subsets of probability measures. Notice that R controls the level of misspecification in the direction that is orthogonal to the line connecting the kernel embeddings of P_X and P_Y. Setting $R = 0$ simply asserts that P_Z is guaranteed to be a mixture of P_X and P_Y, as is the case for prior works on LFHT. Before presenting our main result on the minimax sample complexity of mLFHT, let us define one final piece of terminology. We say that a test Ψ, which takes some data as input and takes values in $\{0, 1\}$, has total error probability less than α for the problem of testing H_0 vs H_1 if

$$\sup_{P \in H_0} P(\Psi = 1) + \sup_{Q \in H_1} Q(\Psi = 0) \leq \alpha. \tag{3.3.2}$$

Theorem 3.3.1. *Suppose we observe three i.i.d. samples X, Y, Z from distributions P_X, P_Y, P_Z composed of n, n, m observations respectively and let $C \in (0, \infty)$ and $R, \epsilon \geq 0$ and $\delta \in (0, 1)$. There exists a universal constant $c > 0$ such that $\Psi_{\delta/2}$ defined Section 3.2.2 tests H_0 vs H_1, as defined in (3.3.1), at total error α provided*

$$\min\{m, n\} \geq c \frac{C\|\lambda\|_\infty \log(1/\alpha)}{(\epsilon\delta/(1+R))^2} \qquad \text{and} \qquad \min\{n, \sqrt{nm}\} \geq c \frac{C\|\lambda\|_2 \log(1/\alpha)}{\delta\epsilon^2}.$$

Note that Theorem 3.3.1 does *not* place assumptions on the distributions P_X, P_Y beyond bounded density with respect to the base measure μ. This is different from usual results in statistics, where prior specification of distribution classes is crucial. On the other hand, instead of standard distances such as L^p, we assume separation with respect to MMD and the latter is potentially harder to interpret than, say, L^1 i.e. total variation. We do point out that our Theorem 3.3.1 can be used to derive results in the classical setting; we discuss this further in Section 3.3.4.

In an appropriate regime of the parameters, the sufficient sample complexity in Theorem 3.3.1 exhibits a trade-off of the form $\min\{n, \sqrt{mn}\} \gtrsim \|\lambda\|_2 \log(1/\alpha)/(\delta\epsilon^2)$ between the number of simulation samples n and real observations m. This trade-off is shown in Figure 3.2 using data from a toy problem. The trade-off is clearly asymmetric and the relationship $m \cdot n \geq const.$ also seems to appear. In this toy problem we set $R = 0, \delta = 1, \epsilon = .3, k = 100$ and $P_X = P_Z, P_Y$ are distributions on $\{1, 2, \ldots, k\}$ with $P_X(i) = (1 + \epsilon \cdot (2 \cdot \mathbb{1}\{i \text{ odd}\} - 1))/k = 2/k - P_Y(i)$ for all $i = 1, 2, \ldots, k$. The kernel we take is $K(x, y) = \sum_{i=1}^{k} \mathbb{1}\{x = y = i\}$ and μ is simply the counting measure; the resulting MMD is simply the L^2-distance on pmfs.

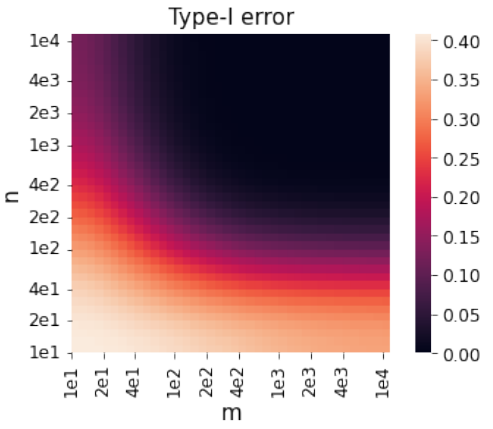


Figure 3.2: n versus m trade-off for the toy experiment, verifying Theorem 3.3.1. Probabilities estimated over 10^4 runs, and smoothed using Gaussian noise.

Figure 3.1 illustrates a larger scale experiment using real data using a trained kernel. Note that we plot the *total* number n of simulation samples, including those used for *training* the kernel itself (see Section 3.4); which ensures that Figure 3.1 gives a realistic picture of data requirements. However, due to the dependence between the kernel and the data, Theorem 3.3.1 no longer applies. Nevertheless, we observe a trade-off similar to Figure 3.2.

3.3.2 Lower Bounds on the Minimax Sample Complexity of (mLFHT)

In this section we prove a minimax lower bound on the sample complexity of mLFHT, giving a partial converse to Theorem 3.3.1. Before we can state this results, we must make some technical definitions. Given $J \geq 2$, let $\|\lambda\|_{2J}^2 := \sum_{j=2}^{J} \lambda_j^2$ and define

$$J_\epsilon^\star := \max\left\{ J : \sup_{\eta_j = \pm 1} \left\| \sum_{j=2}^{J} \eta_j \sqrt{\lambda_j} e_j \right\|_\infty \leq \frac{\|\lambda\|_{2J}}{2\epsilon} \right\}.$$

Theorem 3.3.2 (Lower Bounds for mLFHT). *Suppose that* $\int_{\mathcal{X}} K(x, y)\mu(\mathrm{d}x) \equiv \lambda_1$, $\mu(\mathcal{X}) = 1$ *and* $\sup_{x \in \mathcal{X}} K(x, x) \leq 1$. *There exists a universal constant* $c > 0$ *such that any test of* H_0 *vs* H_1, *as defined in* (3.3.1), *with total error at most* α *must use a number* (n, m) *of observations that satisfy*

$$m \geq c\frac{\lambda_2 \log(1/\alpha)}{\epsilon^2\delta^2} \quad \text{and} \quad n \geq c\frac{\|\lambda\|_{2J_\epsilon^\star}\sqrt{\log(1/\alpha)}}{\epsilon^2} \quad \text{and} \quad \delta m + \sqrt{mn} \geq c\frac{\|\lambda\|_{2J_\epsilon^\star}\sqrt{\log(1/\alpha)}}{\epsilon^2\delta}.$$

Remark 16. *Recall that the eigenvalues* λ *depend on the choice of* μ, *so that by choosing a different base measure* μ *one can optimize the lower bound. However, since* P_X, P_Y, P_Z *are assumed to have bounded density with respect to* μ, *this appears rather involved.*

Remark 17. *The requirements* $\sup_{x \in \mathcal{X}} K(x, x) \leq 1$ *and* $\mu(\mathcal{X}) = 1$ *are is essentially without loss of generality, as* μ *and* K *can be rescaled. The condition* $\int_{\mathcal{X}} K(x, y)\mu(\mathrm{d}x) \equiv \lambda_1$ *implies that the top eigenfunction* e_1 *is equal to a constant or equivalently, that* $y \mapsto K(x, y)\mu(\mathrm{d}x)$ *defines a Markov kernel up to a normalizing constant.*

3.3.3 Tightness of Theorems 3.3.1 and 3.3.2

Dependence on $\|\lambda\|_2$ An apparent weakness of Theorem 3.3.2 is its reliance on the unknown value $J^\star_\epsilon$, which depends on the specifics of the kernel K and base measure μ. Determining it is potentially highly nontrivial even for simple kernels. Slightly weakening Theorem 3.3.2 we obtain the following corollary, which shows that the dependence on $\|\lambda\|_2$ is tight, at least for small ϵ.

Corollary 3.3.3. *Suppose $J \geq 2$ is such that $\sum_{j=2}^{J} \lambda_j^2 \geq c^2\|\lambda\|_2^2$ for some $c \leq 1$. Then $\|\lambda\|_{2J^\star_\epsilon}$ can be replaced by $c\|\lambda\|_2$ in Theorem 3.3.2 whenever $\epsilon \leq c\|\lambda\|_2/(2\sqrt{J-1})$.*

Dependence on R and α Due to the general nature of our lower bound constructions, it is difficult to capture the dependence on the misspecification parameter R. As for the probability of error α, based on related work [62] we expect the gap of size $\sqrt{\log(1/\alpha)}$ to be a shortcoming of Theorem 3.3.1 and not the lower bound. Closing this gap may require a different approach, however, as tests based on empirical L^2 distances are known to have a sub-optimal concentration [102].

Dependence on δ The correct dependence on the signal rate δ is the most important question left open by our theoretical results. Any method requiring n larger than a function of δ irrespective of m (as in Theorem 3.3.1) is provably sub-optimal because taking $m \gtrsim 1/(\delta\epsilon)^2$ and n large enough to estimate both P_X, P_Y to within accuracy $\epsilon/10$ always suffices to reach a fixed level of total error.

3.3.4 Relation to Prior Results

In this section we discuss some connections of Theorem 3.3.1 to prior work. Specifically, we discuss how Theorem 3.3.1 recovers some known results in the literature [10, 139, 78] that are *minimax optimal*. Details omitted in this section are included in Appendix B.2.

Binary Hypothesis Testing Suppose the two distributions P_X, P_Y are *known*, we are given m i.i.d. observations $Z_1, \ldots, Z_m \sim P_Z$ and our task is to decide between the hypotheses $H_0 : P_X = P_Z$ versus $H_1 : P_Y = P_Z$. Then, we may take $n = \infty, R = 0, \delta = 1$ in Theorem 3.3.1 to conclude that

$$m \geq c \cdot \frac{C\|\lambda\|_\infty \log(1/\alpha)}{\epsilon^2}$$

observations suffice to perform the test at total error α.

Two-Sample Testing Suppose we have two i.i.d. samples X and Y, both of size n, from unknown distributions P_X, P_Y respectively and our task is to decide between $H_0 : P_X = P_Y$ against $H_1 : \mathrm{MMD}(P_X, P_Y) \geq \epsilon$. We split our Y sample in half resulting in $Y^{(1)}$ and $Y^{(2)}$ and form the statistic $\Psi_{\mathsf{TS}} := \Psi_{1/2}(X, Y^{(1)}, Y^{(2)}) - \Psi_{1/2}(Y^{(1)}, X, Y^{(2)})$, where $\Psi_{1/2}$ is defined in Section 3.2.2. Then $|\mathbb{E}\Psi_{\mathsf{TS}}|$ is equal to 0 under the null hypothesis and is at least $1 - 2\alpha_1$ under the alternative, where α_1 is the target total error probability of $\Psi_{1/2}$. Taking $\alpha_1 = 5\%$, by repeated sample splitting and majority voting we may amplify the success probability to α provided

$$n \geq c' \frac{C\|\lambda\|_2 \log(1/\alpha)}{\epsilon^2}, \tag{3.3.3}$$

where $c' > 0$ is universal (see Appendix for details). The upper bound (3.3.3) partly recovers [139, Theorem 3 and 5] where authors show that thresholding the MMD with Gaussian

kernel $G_\sigma(x,y) = \sigma^{-d}\exp(-\|x-y\|^2/\sigma^2)$ achieves the minimax optimal sample complexity $n \asymp \epsilon^{-(2\beta+d/2)/\beta}$ for the problem of two-sample testing over the class $\mathcal{P}_{\beta,d}$ of d-dimensional $(\beta,2)$-Sobolev-smooth distributions (defined in Appendix B.2.3) under ϵ-L^2-separation. For this, taking $\sigma \asymp \epsilon^{1/\beta}$ ensures that $\|P-Q\|_{L^2} \lesssim \mathrm{MMD}(P,Q)$ over $P,Q \in \mathcal{P}_{\beta,d}$. Taking e.g. $\mathrm{d}\mu(x) = \exp(-\|x\|_2^2)\mathrm{d}x$ as the base measure, (3.3.3) recovers the claimed sample complexity since $\|\lambda\|_2^2 = \int G_\sigma^2(x,y)\mathrm{d}\mu(x)\mathrm{d}\mu(y) = \mathcal{O}(\sigma^{-d})$ hiding dimension dependent constants. Our result requires a bounded density with respect to a Gaussian.

Likelihood-Free Hypothesis Testing By taking $\alpha \asymp R \asymp \delta = \Theta(1)$ in Theorem 3.3.1 we can recover many of the results in [123, 124, 78]. When $\mathcal{X}$ is finite, we can take the kernel $K(x,y) = \sum_{z\in\mathcal{X}} \mathbb{1}\{x=y=z\}$ in Theorem 3.3.1 to obtain the results for bounded discrete distributions (defined in Appendix B.2.1) which state that under ϵ-TV-separation the minimax optimal sample complexity is given by $m \gtrsim 1/\epsilon^2; \min\{n, \sqrt{nm}\} \gtrsim \sqrt{|\mathcal{X}|}/\epsilon^2$. A similar kernel recovers the optimal result for the class of β-Hölder smooth densities on the hypercube $[0,1]^d$ (see Appendix B.2.2).

Curse of Dimensionality Using the Gaussian kernel G_σ as for two-sample testing above, one can conclude by Theorem 3.3.1 that the required number of samples for (mLFHT) over the class $\mathcal{P}_{\beta,d}$ under ϵ-L^2-separation grows at most like $\Omega\left((\frac{c}{\epsilon})^{\Omega(d)}\frac{1}{\delta^2}\right)$ for some $c>1$, instead of the expected $\Omega\left((\frac{c}{\epsilon\delta})^{\Omega(d)}\right)$. This may be interpreted as theoretical support for the success of LFI in practice where signal and background can be rather different (cf. [14, Figures 2-3]) and the difficulty of the problem stems from the rate of signal events being small (i.e. $\epsilon \approx 1$ but $\delta \ll 1$).

3.4 Learning Kernels from Data

Given a *fixed* kernel K, our Theorems 3.3.1 and 3.3.2 show that the sample complexity is heavily dependent on the separation ϵ under the given MMD as well as the spectrum $\lambda = \lambda(\mu, K)$ of the kernel. Thus, to have good test performance we need to use a kernel K that is well-adapted to the problem at hand. In practice, however, instead of using a fixed kernel it would be only natural to use part of the simulation sample to try to *learn* a good kernel.

In Sections 3.4 and 3.5 we report experimental results after *training* a kernel parameterized by a neural network on part of the simulation data. In particular, due to the dependence between the data and the kernel, Theorem 3.3.1 doesn't directly apply. Our main contribution here is showing the existence of an asymmetric simulation-experimentation trade-off (cf. Figure 3.1 and also Section 3.1.1) even in this realistic setting. Figure 3.1 plots the *total* number n of simulations used, including those used for training, so as to provide a realistic view of the amount of data used. The experiments also illustrate that the (trained-)kernel-based statistic of Section 3.2.2 achieves state-of-the-art performance.

3.4.1 Proposed Training Algorithm

Consider splitting the data into three parts: $(X^{\mathsf{tr}}, Y^{\mathsf{tr}})$ is used for training (optimizing) the kernel; $(X^{\mathsf{ev}}, Y^{\mathsf{ev}})$ is used to evaluate our test statistic at test time; and $(X^{\mathsf{cal}}, Y^{\mathsf{cal}})$ is used for calibrating the distribution of the test statistic under the null hypothesis. We write $n_s = |X^s| = |Y^s|$ for $s \in \{\mathsf{tr}, \mathsf{ev}, \mathsf{cal}\}$. Given the training data $X^{\mathsf{tr}}, Y^{\mathsf{tr}}$ with empirical measures

Algorithm 1 mLFHT with a learned deep kernel

Input: $(X^{\mathsf{tr}}, X^{\mathsf{ev}}, X^{\mathsf{cal}})$, $(Y^{\mathsf{tr}}, Y^{\mathsf{ev}}, Y^{\mathsf{cal}})$; parametrized kernel K_ω; hyperparameters and initialization.
\# **Phase 1:** *Kernel training (optimization) on* X^{tr} *and* Y^{tr}.
$\omega \leftarrow \arg\max_\omega^{\text{Optimizer}} \widehat{J}(X^{\mathsf{tr}}, Y^{\mathsf{tr}}; K_\omega)$; *# maximize objective as defined in (3.4.1)*
\# **Phase 2:** *Distributional calibration of test statistic (under null hypothesis).*
for $r = 1, 2, \ldots, k$ **do**
$Z^{\mathsf{cal},r} \leftarrow$ sample m points without replacement from X^{cal};
$T_r \leftarrow \frac{1}{n_{\mathsf{ev}} m} \sum_{i,j} \left(K_\omega(Z_i^{\mathsf{cal},r}, Y_j^{\mathsf{ev}}) - K_\omega(Z_i^{\mathsf{cal},r}, X_j^{\mathsf{ev}}) \right)$;
end for
\# **Phase 3:** *Inference with input* Z.
$\widehat{T} \leftarrow \frac{1}{n_{\mathsf{ev}} m} \sum_{i,j} \left(K_\omega(Z_i, Y_j^{\mathsf{ev}}) - K_\omega(Z_i, X_j^{\mathsf{ev}}) \right)$;
Output: Estimated p-value: $\frac{1}{k} \sum_{i=1}^k \mathbb{1}\{\widehat{T} < T_i\}$.

$\widehat{P}_{X^{\mathsf{tr}}}, \widehat{P}_{Y^{\mathsf{tr}}}$, we maximize the objective in

$$\widehat{J}(X^{\mathsf{tr}}, Y^{\mathsf{tr}}; K) = \frac{\mathrm{MMD}_u^2(\widehat{P}_{X^{\mathsf{tr}}}, \widehat{P}_{Y^{\mathsf{tr}}}; K)}{\widehat{\sigma}(X^{\mathsf{tr}}, Y^{\mathsf{tr}}; K)}, \tag{3.4.1}$$

which was introduced in [187]. Here $\widehat{\sigma}^2$ is an estimator of the variance of $\mathrm{MMD}_u^2(\widehat{P}_{X^{\mathsf{tr}}}, \widehat{P}_{Y^{\mathsf{tr}}}; K)$ and is defined in Appendix B.6.1.

Intuitively, the objective J aims to separate P_X from P_Y while keeping variance low. For a heuristic justification of its use for (mLFHT) see Appendix.

In Algorithm 1 we describe the training and testing procedure, which produces unbiased p-values for (mLFHT) when there is no misspecification ($R = 0$ in Theorem 3.3.1). During training, we use the Adam optimizer [128] with stochastic batches.

Proposition 3.4.1. *When there is no misspecification ($R = 0$ in Theorem 3.3.1), Algorithm 1 outputs an unbiased estimate of the p-value that is consistent as* $\min\{n_{\mathsf{cal}}, k\} \to \infty$.

Time complexity Algorithm 1 runs in three separate stages: training, calibration, and inference. The first two take $O\left(\#\text{epochs} \cdot B^2 + k n_{\mathsf{ev}} m\right)$ total time, where B is the batch size, whereas Phase 3 takes only $O(n_{\mathsf{ev}} m)$ time, which is generally much faster especially if $n_{\mathsf{ev}} \ll n_{\mathsf{tr}}$.

Sample usage Empirically, data splitting in Algorithm 1 can have non-trivial effects on performance. Instead of training the kernel on only a fraction of the data ($\{X^{\mathsf{tr}}, Y^{\mathsf{tr}}\} \cap \{X^{\mathsf{ev}}, Y^{\mathsf{ev}}\} = \varnothing$), we discovered that taking $\{X^{\mathsf{ev}}, Y^{\mathsf{ev}}\} \subseteq \{X^{\mathsf{tr}}, Y^{\mathsf{tr}}\}$ results in more efficient use of data. The stronger condition $\{X^{\mathsf{ev}}, Y^{\mathsf{ev}}\} = \{X^{\mathsf{tr}}, Y^{\mathsf{tr}}\}$ can also be applied; we take $\subseteq$ to reduce time complexity. We do, however, crucially require $X^{\mathsf{cal}}, Y^{\mathsf{cal}}$ in Phase 2 to be independently sampled ("held-out") for consistent p-value estimation. Finally, we remark also that splitting this way is only valid in the context of Algorithm 1. For the test (3.2.4) using the data-dependent threshold γ, one needs $\{X^{\mathsf{tr}}, Y^{\mathsf{tr}}\} \cap \{X^{\mathsf{ev}}, Y^{\mathsf{ev}}\} = \varnothing$ to estimate γ.

3.4.2 Classifier-Based Tests and Other Benchmarks

Let $\phi : \mathcal{X} \to [0, 1]$ be a classifier, assigning small values to P_X and high values to P_Y by minimizing cross-entropy loss of a classifier net. There are two natural test statistics based on ϕ:

Scheffé's Test. The first idea, attributed to Scheffé in folklore [59, Section 6], is to take the statistic $T(Z) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{\phi(Z_i) > t\}$ where t is some (learn-able) threshold.

Approximate Neyman-Pearson / Logit Methods. If ϕ is trained to perfection, then $\phi(z) = \mathbb{P}(P_X|z)$ would be the likelihood and $\phi(z)/(1-\phi(z))$ would equal precisely the likelihood ratio between P_Y and P_X at z. This motivates the use of $T(Z) = \frac{1}{m}\sum_{i=1}^{m} \log(\phi(Z_i)/(1-\phi(Z_i)))$. See also [42].

Let us list the testing procedures that we benchmark against each other in our experiments.

1. **MMD-M**: The MMD statistic (3.2.3) using K with the mixing architecture

$$K(x,y) = [(1-\tau)G_\sigma(\varphi_\omega(x), \varphi_\omega(y)) + \tau] \cdot G_{\sigma_0}(x + \varphi'_{\omega'}(x), y + \varphi'_{\omega'}(y)).$$

   Here G_σ is the Gaussian kernel with variance σ^2; $\varphi_\omega, \varphi'_{\omega'}$ are NN's (with parameters ω, ω'), and $\sigma, \sigma_0, \tau, \omega, \omega'$ are trained.

2. **MMD-G**: The MMD statistic (3.2.3) using the Gaussian kernel architecture $K(x,y) = G_\sigma(\varphi_\omega(x), \varphi_\omega(y))$ where φ_ω is the feature mapping parametrized by a trained network and σ is a trainable parameter.

3. **MMD-O**: The MMD statistic (3.2.3) using the Gaussian Kernel $K(x,y) = G_\sigma(x,y)$ with optimized bandwidth σ. First proposed in [31, 143].

4. **UME**: An interpretable model comparison algorithm proposed by [118], which evaluates the kernel mean embedding on a chosen "witness set".

5. **SCHE**, **LBI**: Scheffé's test and Logit Based Inference methods [42], based on a binary classifier network ϕ trained via cross-entropy, introduced above.

6. **RFM**: Recursive Feature Machines, a recently proposed kernel learning algorithm by [170].

3.4.3 Additive Statistics and the Thresholding Trick

Given a function f (usually obtained by training) and test data $Z = (Z_1, \ldots, Z_m)$, we call a test additive if its output is obtained by thresholding $T_f(Z) := \frac{1}{m}\sum_{i=1}^{m} f(Z_i)$. We point out that all of **MMD-M/G/O**, **SCHE**, **LBI**, **UME**, **RFM** are of this form, see the Appendix for further details. Similarly to [143], we observe that any such statistic can be realized by our kernel-based approach.

Proposition 3.4.2. *The kernel-based statistic defined in (3.2.3) with the kernel $K(x,y) = f(x)f(y)$ is equal to T_f up to a multiplicative and additive constant independent of Z.*

Motivated by the Scheffé's test, instead of directly thresholding the additive statistic $T_f(Z)$, we found empirically that replacing f by $f_t(x) := \mathbb{1}\{f(x) > t\}$ can yield improved power. We set t by maximizing an estimate of the significance under the null using a normal approximation, i.e. by solving $t_{\mathsf{opt}} := \arg\max_t \frac{T_{f_t}(Y^{\mathsf{opt}}) - T_{f_t}(X^{\mathsf{opt}})}{\sqrt{T_{f_t}(X^{\mathsf{opt}})(1-T_{f_t}(X^{\mathsf{opt}}))}}$, where $X^{\mathsf{opt}}, Y^{\mathsf{opt}}$ satisfy $\{X^{\mathsf{opt}}, Y^{\mathsf{opt}}\} \cap (\{X^{\mathsf{cal}}, Y^{\mathsf{cal}}\} \cup \{X^{\mathsf{ev}}, Y^{\mathsf{ev}}\}) = \varnothing$. This trick improves the performance of our tester on the Higgs dataset in Section 3.5.2 but not for the image detection problem in Section 3.5.1.

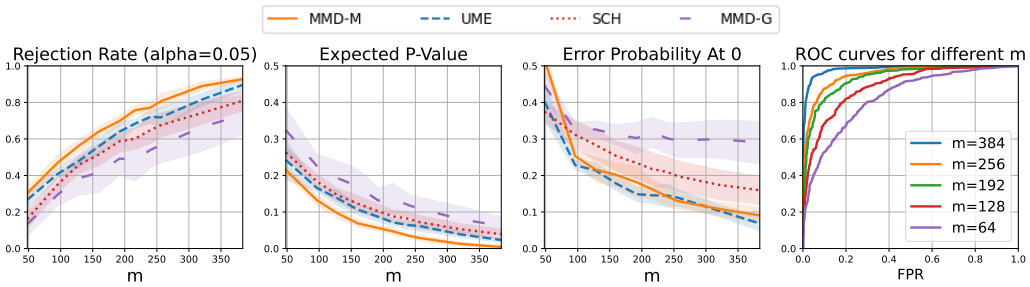


Figure 3.3: Empirical performance on (3.5.1) for the CIFAR detection problem when $n_{\text{tr}} = 1920$. Plots from left to right are as follows. (a) rejection rate under the alternative if test rejects whenever the estimated p-value is smaller than 5%; (b) expected p-value [178] under the alternative; (c) the average of type-I and II error probabilities when thresholded at 0 (different from (3.2.4), see Appendix); and (d) ROC curves for different m using MMD-M and Algorithm 1. Shaded area shows the standard deviation across 10 independent runs. Missing benchmarks (thresholded MMD, MMD-O, LBI, RFM) are weaker; see Appendix for full plot.

3.5 Experiments

Our code can be found at https://github.com/Sr-11/LFI.

3.5.1 Image Source Detection

Our first empirical study looks at the task of detecting whether images come from the CIFAR-10 [136] dataset or a SOTA generative model (DDPM) [95, 167]. While source detection is on its own interesting, it turns out that detecting whether a group of images comes from the generative model versus the real dataset can be too "easy" (see experiments in [118]). Therefore, we consider a *mixed alternative*, where the alternative hypothesis is not simply the generative model but CIFAR with planted DDPM images. Namely, our n labeled images come from the following distributions:

$$P_X = \text{CIFAR}, \quad \text{and} \quad P_Y = \frac{1}{3} \cdot \text{DDPM} + \frac{2}{3} \cdot \text{CIFAR}. \tag{3.5.1}$$

The goal is to test whether the m unlabeled observations Z have been corrupted with ρ or more fraction of DDPM images (versus uncorrupted CIFAR); this corresponds to (LFHT) (or equivalently (mLFHT) with $\delta = 1$). Figure 3.3 shows the performance of our approach with this mixed alternative.

Network Architecture With a standard deep CNN, the difference is only at the final layer: for the kernel-based tests it is a feature output; for classifiers, we add an extra linear layer to logits.

We see from Figure 3.3 that our kernel-based test outperforms other benchmarks at a fixed training set size n_{tr}. One potential cause is that MMD has an "optimization" subroutine (which it solves in closed form) as it is an IPM. This additional layer of optimization may lead to better performance at small sample sizes. The thresholding trick does not seem to improve power empirically. We omit several benchmarks from this figure for graphic presentation and they do not exhibit good separating power; see the Appendix for the complete results. The bottom plot of Figure 3.1 shows m and n on log-scale against the total probability of error, exhibiting the simulation-experimentation trade-off.

3.5.2 Higgs-Boson Discovery

The 2012 announcement of the Higgs boson's discovery by the ATLAS and CMS experiments [1, 40] marked a significant milestone in physics. The statistical problem inherent in the experiment is well-modeled by (mLFHT), using a signal rate δ predicted by theory and misspecification parameter $R = 0$ (as was assumed in the original discovery). We consider our algorithm's power against past studies in the physics literature [14] as measured by the *significance of discovery*. We note an important distinction from Algorithm 1 in this application.

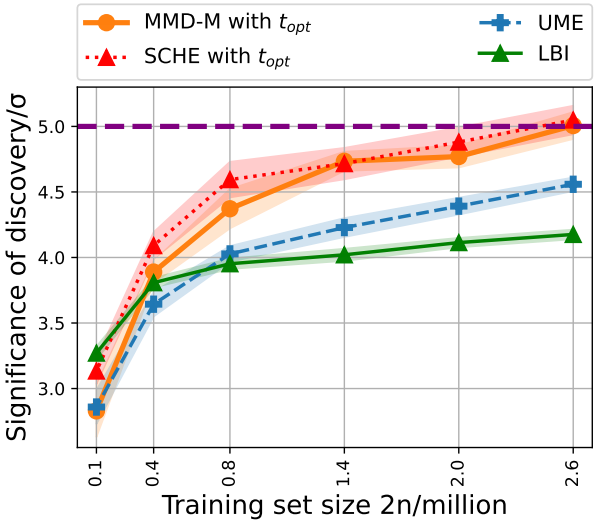


Figure 3.4: Expected significance of discovery on a mixture of 1000 backgrounds and 100 signals in the Higgs experiment. Shaded area shows the standard deviation over 10 independent runs. See Appendix for full plot including missing benchmarks.

Estimating the Significance In physics, the threshold for claiming a "discovery" is usually at a significance of 5σ, corresponding to a p-value of 2.87×10^{-7}. Approximately $n_{\mathsf{cal}} \sim (2.87)^{-1} \times 10^7$ samples would be necessary for Algorithm 1 to reach such a precision. Fortunately the distribution of the test statistic is approximated by a Gaussian customarily. We adopt this approach for our experiment hereby assuming that m is large enough for the CLT to apply. We use the "expected significance of discovery" as our metric [14] which, for the additive statistic $T_f = \frac{1}{m}\sum_{i=1}^{m} f(Z_i)$, is given by $\frac{\delta(T_f(Y^{\mathsf{cal}}) - T_f(X^{\mathsf{cal}}))}{\sqrt{\widehat{\mathrm{var}}(f(X^{\mathsf{cal}}))/m}}$. If the thresholding trick (Section 3.4.3) is applied we use the more precise Binomial tail, in which case the significance is estimated by $-\Phi^{-1}(\mathbb{P}(\mathrm{Bin}(m, T_{f_{t_{\mathsf{opt}}}}(X^{\mathsf{cal}})) \geq T_{f_{t_{\mathsf{opt}}}}(Z)))$, where Φ is the standard normal CDF.

Newtowk Architecture The architecture is a 6-layer feedforward net similar for all tests (kernel-based and classifiers) except for the last layer. We leave further details to the Appendix.

As can be seen in Figure 3.4, Scheffé's test and MMD-M with threshold t_{opt} are the best methods, achieving similar performance as the algorithm of [14]; reaching the significance level of 5σ on 2.6 million simulated datapoints and a test sample made up of a mixture of 1000 backgrounds and 100 signals. The top plot of Figure 3.1 shows m and n on log-scale against the total probability of error through performing the test (3.2.4), exhibiting the asymmetric simulation-experimentation trade-off.

3.6 Conclusion

In this paper, we introduced (mLFHT) as a theoretical model of real-world likelihood-free signal detection problems arising in science. We proposed a kernel-based test statistic and analyzed its minimax sample complexity, obtaining both upper (Theorem 3.3.1) and lower bounds (Theorem 3.3.2) in terms of multiple problem parameters, and discussed their tightness

(Section 3.3.3) and connections to prior work (Section 3.3.4). On the empirical side, we described a method for training a parametrized kernel and proposed a consistent p-value estimate (Algorithm 1 and Proposition 3.4.1). We examined the performance of our method in two experiments and found that parametrized kernels achieve state-of-the-art performance compared to relevant benchmarks from the literature. Moreover, we confirmed experimentally the existence of the asymmetric simulation-experimentation trade-off (Figure 3.1) which is suggested by minimax analysis. We defer further special cases of Theorem 3.3.1, all relevant proofs and experimental details to the Appendix.

Chapter 4

Minimax Optimal Testing via Classification

This chapter is a reproduction of [74], which was published at The Thirty Sixth Annual Conference on Learning Theory, and is joint work with Yanjun Han and Yury Polyanskiy.

4.1 Introduction

The rapid development of machine learning over the past three decades has had a profound impact on many areas of science and technology. It has replaced or enhanced traditional statistical procedures and automated feature extraction and prediction where in the past human experts had to intervene manually. One example is the technique that has become known as 'classification accuracy testing' (CAT). The idea, first explicitly described in [72], is extremely simple. Consider the setting of two-sample testing: suppose the statistician has samples X and Y of size n from two distributions $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ respectively on some space $\mathcal{X}$, and wishes to test the hypotheses

$$H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_{\mathsf{Y}} \qquad \text{versus} \qquad H_1 : \mathbb{P}_{\mathsf{X}} \neq \mathbb{P}_{\mathsf{Y}}. \tag{TS}$$

The statistician has many classical methods at their disposal such as the Kolmogorov-Smirnov or the Wilcoxon – Mann – Whitney test. Friedman's idea was to use machine learning as a powerful tool to summarize the data and subsequently apply a classical two-sample test to the transformed data. More concretely, the proposal is to train a binary classifier $\mathcal{C} : \mathcal{X} \to \{0, 1\}$ on the labeled data $\cup_{i=1}^{n} \{(X_i, 0), (Y_i, 1)\}$ and compare the samples $\mathcal{C}(X_1), \ldots, \mathcal{C}(X_n)$ and $\mathcal{C}(Y_1), \ldots, \mathcal{C}(Y_n)$.

Friedman's idea to use classifiers to summarize data before applying classical statistical analysis downstream can be generalized beyond two-sample testing (TS). Likelihood-free inference (LFI), also known as simulation-based inference (SBI), has seen a flurry of interest recently. In LFI, the scientist has a dataset $Z_1, \ldots, Z_m \overset{\text{iid}}{\sim} \mathbb{P}_{\theta^\star}$ and is given access to a black box simulator which given a parameter θ produces a random variable with distribution $\mathbb{P}_\theta$. The goal is to do inference on $\theta^\star$. The key aspect of the problem, lending the name 'likelihood-free', is that the scientist doesn't know the inner workings of the simulator. In particular its output is not necessarily differentiable with respect to θ and the density of

$\mathbb{P}_\theta$ cannot be evaluated even up to normalization. This setting arises in numerous areas of science where highly complex, mechanistic, stochastic simulators are used such as climate modeling, particle physics, phylogenetics and epidemiology to name a few, and its importance was realized as early as [66]. In this paper we study the problem of likelihood-free hypothesis testing (LFHT) proposed recently in [77] as a simplified model of likelihood-free inference. Compared to two-sample testing, here in addition to the dataset Z of size m, we have two 'simulated' samples X, Y of size n each from $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ respectively. The goal is to test the hypotheses

$$H_0 : Z_i \sim \mathbb{P}_{\mathsf{X}} \qquad \text{versus} \qquad H_1 : Z_i \sim \mathbb{P}_{\mathsf{Y}}. \tag{LFHT}$$

It is important that apriori $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ are only known to belong to a certain ambient (usually non-parametric) class. This stands in contrast with the earliest appearances of (LFHT) in [212, 90], where authors studied the rate of decay of the type-I and type-II error probabilities for fixed $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}$.

In the context of (LFHT) the idea of Friedman materializes as follows. First, train a classifier $\mathcal{C} : \mathcal{X} \to \{0, 1\}$ to distinguish between $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ and second, compare the transformed dataset $\{\mathcal{C}(Z_j))\}_{j=1}^m$ to $\{\mathcal{C}(X_i)\}_{i=1}^n$ and $\{\mathcal{C}(Y_i)\}_{i=1}^n$. The second step compares iid samples of Bernoulli random variables (provided $\mathcal{C}$ is trained on held out data), thus any reasonable test simply thresholds the number of Z_j classified as 1, namely the test is of the form

$$\frac{1}{m}\sum_{j=1}^{m} \mathcal{C}(Z_j) \geq \gamma \tag{4.1.1}$$

for some $\gamma \in [0, 1]$. The idea to classify Z as coming from either $\mathbb{P}_{\mathsf{X}}$ or $\mathbb{P}_{\mathsf{Y}}$ based on the empirical mass on some separating set $S = \mathcal{C}^{-1}(\{1\}) \approx \{\mathrm{d}\mathbb{P}_{\mathsf{Y}}/\mathrm{d}\mathbb{P}_{\mathsf{X}} \geq 1\}$ has been attributed to Scheffé in folklore [60, Section 6]. To illustrate the genuine importance of these ideas, we draw on the famous Higgs boson discovery. In 2012 [40, 5] at the Large Hadron Collider (LHC) a team of physicists announced that they observed the Higgs boson, an elementary particle theorized to exist in 1964. It is regarded as the crowning achievement of the LHC, the most expensive instrument ever built. They achieved this feat via likelihood-free inference, using the ideas of classification accuracy testing/Scheffé's test in particular. As part of their analysis pipeline they trained a boosted decision tree classifier on simulated data and thresholded counts of observations falling in the classification region.

This work was initiated as an attempt to understand the theoretical properties of classifier-accuracy testing, motivated by the clear practical interest in these questions. Our intuition told us that restricting the classifier to have binary output might throw away too much statistical power. In regions with large (small) density ratio, the binary output ought to loose useful information about the (un)certainty of the classifier output. The Neyman-Pearson Lemma phrases this succinctly: the optimal classifier aggregates the log density ratio, while heuristically Scheffé's test aggregates indicators that the log density ratio exceeds some threshold. The operational implication of this would be to train probabilistic classifiers $\mathcal{C} : \mathcal{X} \to \mathbb{R}$ approximating the log density ratio, and to aggregate this $\mathbb{R}$-valued output instead of the binary output. However, our results show that this is not necessary for optimality, at least in the minimax sense.

4.1.1 Informal description of the results

We study the problems of goodness-of-fit testing, two-sample testing and likelihood-free hypothesis testing in a minimax framework (see Section 4.2.1 for precise definitions). Namely, given a family of probability distributions $\mathcal{P}$, we study the minimum number of observations n (and m for LFHT) that are required to perform the test with error probability less than $\delta \in (0, 1/2)$ in the worst case over the distributions $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$. We show for multiple natural classes $\mathcal{P}$ that there exist minimax optimal (with some restrictions) classification accuracy tests.

Let us clarify what we mean by 'classification-accuracy' tests for goodness-of-fit testing (GoF) and the problems TS and LFHT. Suppose we have a sample X of size $2n$ from the unknown distribution $\mathbb{P}_{\mathsf{X}}$. We also have a second sample Y of size $2n$ from $\mathbb{P}_{\mathsf{Y}} \in \mathcal{P}$ which corresponds to the *known* null distribution in the case of GoF and is *unknown* in the case of $\mathsf{TS}, \mathsf{LFHT}$. Finally, for LFHT we have an additional sample Z of size $2m$ from $\mathbb{P}_{\mathsf{Z}} \in \{\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}\}$. Write $\mathcal{D}_{\mathsf{tr}} =: \{X^{\mathsf{tr}}, Y^{\mathsf{tr}}, Z^{\mathsf{tr}}\}$ for the first halves of each sample and $\mathcal{D}_{\mathsf{te}} =: \{X^{\mathsf{te}}, Y^{\mathsf{te}}, Z^{\mathsf{te}}\}$ for the rest. We train a classifier $\mathcal{C} : \mathcal{X} \to \{0, 1\}$ on the input $\mathcal{D}_{\mathsf{tr}}$ that aims to assign 1 to $\mathbb{P}_{\mathsf{X}}$ and 0 to $\mathbb{P}_{\mathsf{Y}}$. Going forward, it will be easier to think of $\mathcal{C}$ in terms of the 'separating set' $S =: \mathcal{C}^{-1}(\{1\})$. Thus, S is a random subset of $\mathcal{X}$ whose randomness comes from $\mathcal{D}_{\mathsf{tr}}$ and potentially an external seed. Given two datasets $\{A_i\}_{i=1}^a, \{B_j\}_{j=1}^b$, we define the classifier-accuracy statistic

$$T_S(A, B) =: \frac{1}{a}\sum_{i=1}^{a} \mathbb{1}\{A_i \in S\} - \frac{1}{b}\sum_{j=1}^{b} \mathbb{1}\{B_j \in S\}. \tag{4.1.2}$$

The name 'classifier-accuracy' is given due to the fact that $T_S(X^{\mathsf{te}}, Y^{\mathsf{te}}) + 1$ is equal to the sum of the fraction of correctly classified test instances under the two classes. Finally, we say a test is a classifier-accuracy test if its output is obtained by thresholding $|T_S|$ for some classifier $\mathcal{C} = \mathbb{1}_S$ on the test data $\mathcal{D}_{\mathsf{te}}$.

Theorem 4.1.1 (informal)**.** *There exist classifier-accuracy tests with minimax (near-)optimal sample complexity for all problems* $\mathsf{GoF}, \mathsf{TS}, \mathsf{LFHT}$ *and multiple classes of distributions* $\mathcal{P}$*.*

4.1.2 Proof sketch

The bulk of the technical difficulty lies in finding a good separating set $S \subseteq \mathcal{X}$. But how do we measure the quality of S? Define the "separation" $\mathrm{sep}(S) =: \mathbb{P}_{\mathsf{X}}(S) - \mathbb{P}_{\mathsf{Y}}(S)$, and the "size" $\tau(S) =: \min\{\mathbb{P}_{\mathsf{X}}(S)\mathbb{P}_{\mathsf{X}}(S^{\mathsf{c}}), \mathbb{P}_{\mathsf{Y}}(S)\mathbb{P}_{\mathsf{Y}}(S^{\mathsf{c}})\}$. The following lemma describes the performance of classifier-accuracy tests (4.1.2) in terms of sep and τ.

Lemma 4.1.2. *Consider the hypothesis testing problem* $H_0 : p = q$ *versus an arbitrary alternative* H_1*. Suppose that the learner has constructed a separating set* S *such that* $|\mathrm{sep}(S)| = |p(S) - q(S)| \geq \underline{\mathrm{sep}}$ *for every* $(p, q) \in H_1$*, and* $\tau(S) = (p(S)(1 - p(S)) \wedge (q(S)(1 - q(S))) \leq \overline{\tau}$ *for every* $(p, q) \in H_0 \cup H_1$*. Then using only the knowledge of* $\overline{\tau}$*, the classifier-accuracy test* (4.1.2) *with* n *test samples from both* p *and* q *and an appropriate threshold achieves type-I and type-II errors at most* δ*, provided that*

$$n \geq c\frac{\log(1/\delta)}{\underline{\mathrm{sep}}}\left(1 + \frac{\overline{\tau}}{\underline{\mathrm{sep}}}\right)$$

for a large enough universal constant $c > 0$.

With Lemma 4.1.2 in hand it is clear how we need to design S. It should satisfy

$$\big|\operatorname{sep}(S)\big| \text{ is big under } H_1, \text{ and } \tau(S) \text{ is small under both } H_0 \text{ and } H_1 \tag{4.1.3}$$

with probability $1-\delta$. The latter condition, namely that τ is small i.e. $\mathcal{C} = \mathbb{1}_S$ is imbalanced, may seem unintuitive as given any two (sufficiently regular) probability distributions there always exists a balanced classifier whose separation is optimal up to constant.

Proposition 4.1.3. *Let $\mathbb{P}, \mathbb{Q}$ be two distributions on a generic probability space $(\mathcal{X}, \mathcal{F})$. Then*

$$\mathsf{TV}(\mathbb{P}, \mathbb{Q}) \le 2 \sup\{\mathbb{P}(\mathcal{C}(X) = 0) - \mathbb{Q}(\mathcal{C}(X) = 0) : \mathbb{P}(\mathcal{C}(X) = 0) = \mathbb{Q}(\mathcal{C}(X) = 1)\},$$

where $\mathcal{C} : \mathcal{X} \to \{0, 1\}$ is a possibly randomized classifier. Here the constant 2 is tight.

Despite Proposition 4.1.3, we find that choosing a highly imbalanced classifier $\mathcal{C}$ is crucial in obtaining the minimax sample complexity in some classes. This has interesting implications for practical classifier-accuracy testing. Indeed, classifiers are commonly trained to minimize some proxy of misclassification error; however, the above heuristics show that this is not necessarily optimal, instead one should seek *imbalanced* classifiers with large separation. Another way to phrase it is that when training a classifier for testing one should have the downstream task in mind, namely, maximizing the power of the resulting test, and not classification accuracy.

4.1.3 Prior work and contribution

The problem of two-sample (TS) testing (aka closeness testing) and the related problem of goodness-of-fit (GoF) testing (aka identity testing) has a long history in both statistics and computer science. We only mention a small subset of the literature, directly relevant to our work. In seminal works Ingster studied (GoF) for the Gaussian sequence model [112, 113] and for smooth densities [110] in one dimension. Extensions to multiple dimensions and (TS) can be found in works such as [139, 10]. For discrete distributions on a large alphabet the two problems appeared first in [81, 20], see also [39, 199] and the survey [36]. Recent work [63, 62] has focused on GoF and TS with vanishing error probability.

The problem of likelihood-free hypothesis testing appeared first in the works [212, 90], who studied the asymptotic setting. Minimax likelihood-free hypothesis testing (LFHT) was first studied by the information theory community in [123, 124] for a restricted class of discrete distributions on a large alphabet, with a strengthening by [102] to vanishing error probability (in some regimes). More recently, the problem was proposed in [77] as a simplified model of likelihood-free inference, and authors derived minimax optimal sample complexities for constant error in the settings studied in the present paper.

The idea of using classifiers for two-sample testing was proposed in [72] and has seen a flurry of interest [83, 144, 126, 94]. In likelihood-free inference the output of classifiers can be used as summary statistics for Approximate Bayesian Computation [117, 91] or to approximate density ratios [51] via the 'likelihood-ratio trick'. A classifier with binary $\{0, 1\}$ output was used in the discovery of the Higgs boson [40, 5] to determine the detection region.

Our work is the first to study the non-asymptotic properties of classifier-based tests in any setting and we find that classifier-accuracy tests are minimax optimal for a wide range of problems. As a consequence of our results we resolve the minimax high probability sample complexity of LFHT over all classes studied, and also obtain new, tight results on high probability GoF and TS.

4.1.4 Structure

In Sections 4.2.1 and 4.2.1 we define the statistical problems and distribution classes we study. In Tables 4.1 and 4.2 we present all sample complexity results, and in Section 4.2.2 we indicate how to derive them. Sections 4.3.1, 4.3.2 and 4.3.3 study the problem of learning good separating sets for discrete and smooth distributions and the Gaussian sequence model respectively. The appendix contains all proofs omitted from the main text, including all lower bounds in Appendix C.4.

4.2 Results

4.2.1 Technical preliminaries

Two-sample, goodness-of-fit and likelihood-free hypothesis testing

Formally, we define a hypothesis as a set of probability measures. Given two hypotheses H_0 and H_1 consisting of distributions on some measurable space $\mathcal{X}$, we say that a function $\psi : \mathcal{X} \to \{0, 1\}$ tests the two hypotheses against each other with error at most $\delta \in (0, 1/2)$ if

$$\max_{i=0,1} \max_{P \in H_i} \mathbb{P}_{S \sim P}(\psi(S) \neq i) \leq \delta. \tag{4.2.1}$$

Throughout the remainder of this section let $\mathcal{P}$ be a class of probability distributions on $\mathcal{X}$. Suppose we observe independent samples $X \sim \mathbb{P}_{\mathsf{X}}^{\otimes n}$, $Y \sim \mathbb{P}_{\mathsf{Y}}^{\otimes n}$ and $Z \sim \mathbb{P}_{\mathsf{Z}}^{\otimes m}$ whose distributions $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Z}} \in \mathcal{P}$ are *unknown* to us. We now define the problems at the center of our work.

Definition 6. *Given a known $\mathbb{P}_0 \in \mathcal{P}$, **goodness-of-fit testing** is the comparison of*

$$H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_0 \qquad \textit{versus} \qquad H_1 : \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_0) \geq \epsilon \tag{GoF}$$

based on the sample X. Write $n_{\mathsf{GoF}}(\epsilon, \delta, \mathcal{P})$ for the smallest number such that for all $n \geq n_{\mathsf{TS}}$ there exists a function $\psi : \mathcal{X}^n \to \{0, 1\}$ which given X as input tests between H_0 and H_1 with error probability at most δ, for arbitrary $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_0 \in \mathcal{P}$.

Definition 7. ***Two-sample testing** is the comparison of*

$$H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_{\mathsf{Y}} \qquad \textit{versus} \qquad H_1 : \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \geq \epsilon \tag{TS}$$

based on the samples X, Y. Write $n_{\mathsf{TS}}(\epsilon, \delta, \mathcal{P})$ for the smallest number such that for all $n \geq n_{\mathsf{TS}}$ there exists a function $\psi : \mathcal{X}^n \times \mathcal{X}^n \to \{0, 1\}$ which given X, Y as input tests between H_0 and H_1 with error probability at most δ, for arbitrary $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P}$.

Definition 8. ***Likelihood-free hypothesis testing*** *is the comparison of*

$$H_0 : \mathbb{P}_{\mathsf{Z}} = \mathbb{P}_{\mathsf{X}} \qquad \textit{versus} \qquad H_1 : \mathbb{P}_{\mathsf{Z}} = \mathbb{P}_{\mathsf{Y}} \tag{LF}$$

based on the samples X, Y, Z. Write $\mathcal{R}_{\mathsf{LF}}(\epsilon, \delta, \mathcal{P}) \subseteq \mathbb{R}^2$ for the maximal set such that for all $(n, m) \in \mathbb{N}^2$ with $n \geq x, m \geq y$ for some $(x, y) \in \mathcal{R}_{\mathsf{LF}}$, there exists a function $\psi : \mathcal{X}^n \times \mathcal{X}^n \times \mathcal{X}^m \to \{0, 1\}$ which given X, Y, Z as input, successfully tests H_0 against H_1 with error probability at most δ, provided $\mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \geq \epsilon$ and $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P}$.

Classes of distributions

We consider the following nonparametric families of distributions.

Smooth density. Let $\mathcal{C}(\beta, d, C)$ denote the set of functions $f : [0, 1]^d \to \mathbb{R}$ that are $\lceil \beta - 1 \rceil$-times differentiable and satisfy

$$\|f\|_{\mathcal{C}_\beta} =: \max\left(\max_{0 \leq |\alpha| \leq \lceil \beta - 1 \rceil} \|f^{(\alpha)}\|_\infty, \sup_{x \neq y \in [0,1]^d, |\alpha| = \lceil \beta - 1 \rceil} \frac{|f^{(\alpha)}(x) - f^{(\alpha)}(y)|}{\|x - y\|_2^{\beta - \lceil \beta - 1 \rceil}} \right) \leq C,$$

where $\lceil \beta - 1 \rceil$ denotes the largest integer strictly smaller than β and $|\alpha| = \sum_{i=1}^d \alpha_i$ for the multiindex $\alpha \in \mathbb{N}^d$. We write $\mathcal{P}_{\mathsf{H}}(\beta, d, C_{\mathsf{H}})$ for the class of distributions with Lebesgue-densities in $\mathcal{C}(\beta, d, C_{\mathsf{H}})$.

Distributions on a finite alphabet. For $k \in \mathbb{N}$, let

$$\begin{aligned} \mathcal{P}_{\mathsf{D}}(k) &=: \{\text{all distributions on the finite alphabet } [k]\}, \\ \mathcal{P}_{\mathsf{Db}}(k, C_{\mathsf{Db}}) &=: \{p \in \mathcal{P}_{\mathsf{D}}(k) : \|p\|_\infty \leq C_{\mathsf{Db}}/k\}, \end{aligned}$$

where $C_{\mathsf{Db}} > 1$ is a constant. In other words, $\mathcal{P}_{\mathsf{Db}}$ are those discrete distributions that are bounded by a constant multiple of the uniform distribution.

Gaussian sequence model on the Sobolev ellipsoid. Define the Sobolev ellipsoid $\mathcal{E}(s, C)$ of smoothness $s > 0$ and size $C > 0$ as $\{\theta \in \mathbb{R}^{\mathbb{N}} : \sum_{j=1}^\infty j^{2s}\theta_j^2 \leq C\}$. For $\theta \in \mathbb{R}^\infty$ let $\mu_\theta = \otimes_{i=1}^\infty \mathcal{N}(\theta_i, 1)$, and define our second class as

$$\mathcal{P}_{\mathsf{G}}(s, C_{\mathsf{G}}) =: \{\mu_\theta \, : \theta \in \mathcal{E}(s, C_{\mathsf{G}})\}.$$

To briefly motivate the study of $\mathcal{P}_{\mathsf{G}}$, consider the classical Gaussian white noise model. Here we have iid observations of the stochastic process

$$\mathrm{d}Y_t = f(t)\mathrm{d}t + \mathrm{d}W_t, \ t \in [0, 1],$$

where $(W_t)_{t \geq 0}$ denotes Brownian motion and $f \in L^2[0, 1]$ is unknown. Suppose now that $\{\phi_i\}_{i \geq 1}$ forms an orthonormal basis for $L^2[0, 1]$ and given an observation Y define the values

$$y_i =: \langle Y, \phi_i \rangle = \int_0^1 f(t)\phi_i(t)\mathrm{d}t + \int_0^1 \phi_i(t)\mathrm{d}W_t =: \theta_i + \epsilon_i.$$

Notice that $\epsilon_i \sim \mathcal{N}(0, 1)$ and that $\mathbb{E}[\epsilon_i \epsilon_j] = \mathbb{1}_{i=j}$. In other words, the sequence $\{y_i\}_{i \geq 1}$ is an observation from the distribution μ_θ. Consider the particular case of $\phi_1 \equiv 1$ and $\phi_{2k} = \sqrt{2}\cos(2\pi k x), \phi_{2k+1} = \sqrt{2}\sin(2\pi k x)$ for $k \geq 1$ and assume that f satisfies periodic boundary conditions. Then θ denotes the Fourier coefficients of f and the condition that $\sum_{j=1}^\infty j^{2s}\theta_j^2 \leq C$ is equivalent to an upper bound on the order $(s, 2)$-Sobolev norm of f, see e.g. Proposition 1.14 of [198]. In other words, by studying the class $\mathcal{P}_{\mathsf{G}}$ we can deduce results for signal detection in Gaussian white noise, where the signal has bounded Sobolev norm.

Table 4.1: Minimax sample complexity of testing (up to constant factors) over $\mathcal{P}_{\mathsf{H}}, \mathcal{P}_{\mathsf{G}}, \mathcal{P}_{\mathsf{Db}}$.

	n_{GoF}	n_{TS}	$\mathcal{R}_{\mathsf{LF}}$
$\mathcal{P}_{\mathsf{Db}}(k)$	$\frac{\sqrt{k\log(1/\delta)}}{\epsilon^2} + \frac{\log(1/\delta)}{\epsilon^2}$	n_{GoF}	$m \geq \frac{\log(1/\delta)}{\epsilon^2}$ and $n \geq n_{\mathsf{GoF}}$ and $nm \geq n_{\mathsf{GoF}}^2$
$\mathcal{P}_{\mathsf{H}}(\beta, d)$	$\frac{\sqrt{\log(1/\delta)}}{\epsilon^{(2\beta+d/2)/\beta}} + \frac{\log(1/\delta)}{\epsilon^2}$	n_{GoF}	$m \geq \frac{\log(1/\delta)}{\epsilon^2}$ and $n \geq n_{\mathsf{GoF}}$ and $nm \geq n_{\mathsf{GoF}}^2$
$\mathcal{P}_{\mathsf{G}}(s)$	$\frac{\sqrt{\log(1/\delta)}}{\epsilon^{(2s+1/2)/s}} + \frac{\log(1/\delta)}{\epsilon^2}$	n_{GoF}	$m \geq \frac{\log(1/\delta)}{\epsilon^2}$ and $n \geq n_{\mathsf{GoF}}$ and $nm \geq n_{\mathsf{GoF}}^2$

Table 4.2: Minimax sample complexity of testing (up to constant factors) over $\mathcal{P}_{\mathsf{D}}$.

	$n_{\mathsf{GoF}}(\mathcal{P}_{\mathsf{D}})$	$n_{\mathsf{TS}}(\mathcal{P}_{\mathsf{D}})$	$\mathcal{R}_{\mathsf{LF}}(\mathcal{P}_{\mathsf{D}})$	
$k \geq \frac{\log(\frac{1}{\delta})}{\epsilon^4}$	(OPT) $n_{\mathsf{GoF}}(\mathcal{P}_{\mathsf{Db}})$	$\left(\frac{k^2\log(\frac{1}{\delta})}{\epsilon^4}\right)^{\frac{1}{3}}$	$n \geq m$	$m \geq \frac{\log(1/\delta)}{\epsilon^2}$ and $mn^2 \geq kn_{\mathsf{GoF}}^2$
				(OPT) $mn^2 \geq kn_{\mathsf{GoF}}^2$ and $n \geq n_{\mathsf{GoF}}$
	(CAT) $n_{\mathsf{GoF}}\left(\frac{\epsilon}{\log(k)}, \frac{\delta}{k}, \mathcal{P}_{\mathsf{Db}}\right)$		$m > n$	(CAT) $\frac{mn^2}{\log(\frac{k}{\delta})} \geq kn_{\mathsf{GoF}}^2\left(\frac{\epsilon}{\log(k)}, \frac{\delta}{k}\right)$ and $n \geq n_{\mathsf{GoF}}(\frac{\epsilon}{\log(k)}, \frac{\delta}{k})$
$k < \frac{\log(\frac{1}{\delta})}{\epsilon^4}$	$n_{\mathsf{GoF}}(\mathcal{P}_{\mathsf{Db}})$	$n_{\mathsf{GoF}}(\mathcal{P}_{\mathsf{Db}})$	$m \geq \frac{\log(1/\delta)}{\epsilon^2}$ and $n \geq n_{\mathsf{GoF}}$ and $nm \geq n_{\mathsf{GoF}}^2$	

4.2.2 Minimax sample complexity of classifier-accuracy tests

In Tables 4.1 and 4.2 we present our and prior results on the minimax sample complexity of GoF, TS and LFHT; here

- unmarked entries denote minimax optimal results achievable by a classifier-accuracy test;
- entries marked with (OPT) denote minimax optimal results that are not known to be achievable by any classifier-accuracy test;
- entries marked with (CAT) denote the best known result using a classifier-accuracy test.

In the constant error regime ($\delta = \Theta(1)$) the results of Tables 4.1 and 4.2 are well known; for instance, the sample complexities of GoF, TS, and LFHT under $\mathcal{P}_{\mathsf{D}}$ were characterized in [162, 25, 77], respectively[1]. Less is known under the high-probability regime ($\delta = o(1)$): for $\mathcal{P}_{\mathsf{D}}$, n_{GoF} was characterized in [103, 63] for uniformity testing, with the general case following from the flattening reduction [64]; n_{TS} was characterized in [62]. For $\mathcal{R}_{\mathsf{LF}}$, the $k > n$ case for $\mathcal{P}_{\mathsf{Db}}$ is resolved by [102], and the achievability direction of the case $m > n$ of $\mathcal{R}_{\mathsf{LF}}$ for $\mathcal{P}_{\mathsf{D}}$ can be deduced from [62] via the natural reduction between TS and LFHT (see [77]). The

[1] [77] only resolved the minimax sample complexity of LFHT for $\mathcal{P}_{\mathsf{D}}$ up to $\log(k)$-factors in some regimes. However, by combining the classifier accuracy tests of this paper for $m \leq n$ and the reduction to two-sample testing with unequal sample size [25, 62] for $m > n$ these gaps are filled.

remaining upper bounds are achievable by the classifier-accuracy tests below, and the proofs of all lower bounds are deferred to Appendix C.4.

As for the efficacy of classifier-accuracy tests, the upper bounds in Tables 4.1 and 4.2 follow from the combination of Lemma 4.1.2 and the following results:

- $\mathcal{P}_{\mathsf{Db}}$: see Corollary 4.3.4;
- $\mathcal{P}_{\mathsf{H}}$: see Section 4.3.2 and Corollary 4.3.4;
- $\mathcal{P}_{\mathsf{G}}$: see Proposition 4.3.8;
- $\mathcal{P}_{\mathsf{D}}$: for GoF, see Proposition 4.3.1 if $k < \log(1/\delta)/\epsilon^4$, and Proposition 4.3.6 otherwise; for TS, see Proposition 4.3.1; for LFHT, see Proposition 4.3.1 if $n \geq k \wedge m$, and Section 4.3.1 and Proposition 4.3.6 otherwise.

4.3 Learning separating sets

In this section, we construct the separating sets S used in the classifier-accuracy test (4.1.2). Section 4.3.1 is devoted to discrete distribution models $\mathcal{P}_{\mathsf{Db}}$ and $\mathcal{P}_{\mathsf{D}}$, where we need a delicate tradeoff between the expected separation and the size of S. A similar construction in the Gaussian sequence model $\mathcal{P}_{\mathsf{G}}$ is presented in Section 4.3.3.

4.3.1 The discrete case

Given two iid samples X, Y of sizes $N_X, N_Y \overset{iid}{\sim} \mathrm{Poi}(n)$ from unknown discrete distributions $p = (p_1, \ldots, p_k), q = (q_1, \ldots, q_k)$ over a finite alphabet $[k] = \{1, 2, \ldots, k\}$, can we learn a set $\hat{S} \subseteq [k]$ using X, Y that separates p from q? To measure the quality of a given separating set $A \subseteq [k]$, we define two quantities $\mathrm{sep}(A) =: p(A) - q(A)$ and $\tau(A) =: \min\{p(A)p(A^{\mathsf{c}}), q(A)q(A^{\mathsf{c}})\}$. Intuitively, the first quantity $\mathrm{sep}(A)$ measures the separation of A, and the second quantity $\tau(A)$ measures the size of A. Recall that by Lemma 4.1.2, in order to perform the classifier-accuracy test (4.1.2), we aim to find a separating set $\hat{S}$ such that

$$|\mathrm{sep}(\hat{S})| \text{ is large and } \tau(\hat{S}) \text{ is small.} \tag{4.3.1}$$

The rest of this section is devoted to the construction of $\hat{S}$ satisfying (4.3.1), and we will present our results on learning separating sets in order of increasing complexity.

Notation: for a random variable X we write $\sigma^2(X)$ for the optimal sub-Gaussian variance proxy of X. In other words, $\sigma^2(X)$ is the smallest value such that $\mathbb{E}\exp(\lambda(X - \mathbb{E}X)) \leq \exp(\lambda^2\sigma^2(X)/2)$ holds for all $\lambda \in \mathbb{R}$.

A natural separating set

Let $\{X_i, Y_i\}_{i\in[k]}$ be the empirical frequencies of each bin $i \in [k]$ in our samples X, Y, i.e. $nX_i \sim \mathrm{Poi}(np_i)$ and $nY_i \sim \mathrm{Poi}(nq_i)$. A natural separating set is the following:

$$\hat{S}_{1/2} =: \{i : X_i > Y_i \text{ or } X_i = Y_i \text{ and } C_i = 1\},$$

where $C_1, C_2 \dots C_k$ are iid Ber(1/2) random variables. We use the subscript "1/2" to illustrate our tie-breaking rule: when $X_i = Y_i$, the symbol i is added to the set with probability 1/2.

Our first result concerns the separating power of the above set.

Proposition 4.3.1. *Suppose $p, q \in \mathcal{P}_{\mathsf{D}}(k)$ with $\mathsf{TV}(p, q) \geq \epsilon$. There exists a universal constant $c > 0$ such that*

$$\mathbb{P}\left(\mathrm{sep}(\hat{S}_{1/2}) \geq c\epsilon^2 \left(\frac{n}{k} \wedge \sqrt{\frac{n}{k}} \wedge \frac{1}{\epsilon}\right)\right) \geq 1 - \delta,$$

provided $n \geq \frac{1}{c} n_{\mathsf{TS}}(\epsilon, \delta, \mathcal{P}_{\mathsf{D}}(k))$.

Together with the trivial upper bound $\tau(\hat{S}_{1/2}) \leq 1/4$, Proposition 4.3.1 and Lemma 4.1.2 imply that using $\hat{S}_{1/2}$ achieves the minimax sample complexity for the following problems:

- GoF in $\mathcal{P}_{\mathsf{Db}}$ and $\mathcal{P}_{\mathsf{D}}$ as long as $k = \mathcal{O}(\log(1/\delta)/\epsilon^4)$;
- TS in $\mathcal{P}_{\mathsf{Db}}$ as long as $k = \mathcal{O}(\log(1/\delta)/\epsilon^4)$, and in $\mathcal{P}_{\mathsf{D}}$ for all (k, ϵ, δ);
- LFHT in $\mathcal{P}_{\mathsf{Db}}$ as long as $k = \mathcal{O}(\log(1/\delta)/\epsilon^4)$, and in $\mathcal{P}_{\mathsf{D}}$ as long as $n \geq m$.

However, in the remaining regimes the above test could be strictly sub-optimal. This failure comes down to two issues. First, Proposition 4.3.1 requires $n \gtrsim n_{\mathsf{TS}}(\epsilon, \delta, \mathcal{P}_{\mathsf{D}}(k))$ in order to find a good separating set, which can be sub-optimal when the optimal sample complexity for the original testing problem is only $n \gtrsim n_{\mathsf{GoF}}(\epsilon, \delta, \mathcal{P}_{\mathsf{D}}(k))$. Second, the quantity $\tau(\hat{S}_{1/2})$ is $\Omega(1)$ in the general case because the tie-breaking rule adds too many symbols to the set. These issues will be addressed separately in the next two sections.

The "better of two" separating sets

This section aims to find a separating set $\hat{S}$ with essentially the same separation as $\hat{S}_{1/2}$ in Proposition 4.3.1, but with a smaller $\tau(\hat{S})$. The central idea is to use a different tie-breaking rule from $\hat{S}_{1/2}$. Given a subset $D \subseteq [k]$, we define the imbalanced separating sets

$$\hat{S}_{>}(D) = \{i \in D : X_i > Y_i\},$$
$$\hat{S}_{<}(D) = \{i \in D : X_i < Y_i\}.$$

In other words, in both $\hat{S}_{>}$ and $\hat{S}_{<}$, we do not include the symbols with $X_i = Y_i$ in the separating set. Consequently, $|\hat{S}_{>}(D)| \vee |\hat{S}_{<}(D)|$ is upper bounded by the sample size; if in addition q_i is bounded from above uniformly over $i \in D$, this will yield good control of τ for both separating sets $\hat{S}_{>}(D)$ and $\hat{S}_{<}(D)$. In particular, $\tau(\hat{S}_{>}(D)) \vee \tau(\hat{S}_{<}(D)) = \mathcal{O}(1 \wedge (n \max_{i \in D} q_i))$.

Next we aim to show that the above sets achieve good separation. However, there is a subtlety here: removing the ties from $\hat{S}_{1/2}$ may no longer guarantee the desired separation, as illustrated in the following proposition.

Proposition 4.3.2. *Consider the distributions p, q on $[3k]$ with $p_i = \mathbb{1}\{i \le k\}/(2k) + \mathbb{1}\{i > k\}/(4k)$ and $q_i = \mathbb{1}\{i \le k\}/k$. Then, for $n \le 0.6k$,*

$$\mathbb{E}\,\mathrm{sep}(\hat{S}_>([3k])) < 0.$$

Proposition 4.3.2 shows that sticking to only one set $\hat{S}_>$ or $\hat{S}_<$ fails to give the same separation guarantees as Proposition 4.3.1. A priori it may seem that $\hat{S}_>$ is designed to capture elements of the support where p is greater than q, but it fails to do so spectacularly. An intuitive explanation of this phenomenon is as follows. Since the probability of each bin is small ($\lesssim 1/k$) under both p and q, in the small n regime[2] can expect that (a) each bin appears either once or not at all and (b) there is no overlap between the observed bins in sample X and Y. In this heuristic picture, the set $\hat{S}_>$ is simply the set of observed bins in the X-sample. Each X-sample falling in the first k bins contributes $-\frac{1}{2k}$ to the separation, while each X-sample in the last $2k$ bins contributes only $+\frac{1}{4k}$ to the separation. Since p puts mass $1/2$ on both the first k and last $2k$ bins, there is an equal number of $n/2$ observations in each part and the overall separation is $\asymp -\frac{n}{8k}$. Similar results can be proved for $\hat{S}_<$ with p, q as above but swapped, and also for modified p, q separated by smaller ϵ in TV for any $\epsilon \in (0, 1)$.

Motivated by the above discussion, in the sequel we consider the sets $\hat{S}_>, \hat{S}_<$ jointly. Specifically, the next proposition shows that *at least one of* the sets $\hat{S}_>$ and $\hat{S}_<$ have a good separation.

Proposition 4.3.3. *There exists a universal constant $c > 0$ such that for any $D \subseteq [k]$ and probability mass functions p, q, it holds that*

$$\mathbb{E}\left[\mathrm{sep}(\hat{S}_>(D)) - \mathrm{sep}(\hat{S}_<(D))\right] \ge c \sum_{i \in D} \frac{n(p_i - q_i)^2}{\sqrt{n(p_i \wedge q_i) + 1}} \wedge |p_i - q_i|,$$
$$\sigma^2(\mathrm{sep}(\hat{S}_>(D))) + \sigma^2(\mathrm{sep}(\hat{S}_<(D))) \le \frac{1}{c} \sum_{i \in D} \frac{p_i + q_i}{n} \wedge |p_i - q_i|^2.$$

Based on Proposition 4.3.3, our final separating set is chosen from these two options, based on evaluation on held out data. As for the choice of D, in this section we choose $D = [k]$. The following corollary summarizes the performance of this choice under $\mathcal{P}_{\mathsf{Db}}$.

Corollary 4.3.4. *Suppose $p, q \in \mathcal{P}_{\mathsf{Db}}(k, \mathcal{O}(1))$ with $\mathsf{TV}(p, q) \ge \epsilon$. There exists a universal constant $c > 0$ such that using the samples X, Y we can find a set $\hat{S} \subseteq [k]$ which, with probability $1 - \delta$, satisfies*

$$\left|\mathrm{sep}(\hat{S})\right| \ge c\epsilon^2 \left(\frac{1}{\epsilon} \wedge \sqrt{\frac{n}{k}} \wedge \frac{n}{k}\right) \qquad \text{and} \qquad \tau(\hat{S}) \le \frac{1}{c}\left(1 \wedge \frac{n}{k}\right), \tag{4.3.2}$$

provided $n \ge \frac{1}{c} n_{\mathsf{GoF}}(\epsilon, \delta, \mathcal{P}_{\mathsf{Db}}(k, \mathcal{O}(1)))$.

By Corollary 4.3.4 and Lemma 4.1.2, using the above set $\hat{S}$ achieves the minimax sample complexity for all problems GoF, TS, and LFHT and all parameters (k, ϵ, δ) under $\mathcal{P}_{\mathsf{Db}}$.

[2]Technically, to satisfy the stated conditions we would require $n \lesssim \sqrt{k}$, but the described event captures dominant effects even for larger $\sqrt{k} \ll n \ll k$.

However, under $\mathcal{P}_{\mathsf{D}}$, the performance of $\hat{S}$ is no better than that of $\hat{S}_{1/2}$. This is because a good control of $\tau(\hat{S}_{>}([k]))$ requires a bounded probability mass function; in other words, choosing $D = [k]$ is not optimal for finding the best separating set under $\mathcal{P}_{\mathsf{D}}$. In the next section, we address this issue by choosing D to be one of $\mathcal{O}(\log k)$ subsets of $[k]$.

The "best of $\mathcal{O}(\log k)$" separating sets

This section is devoted to the two missing regimes $m \geq n$ for LFHT over $\mathcal{P}_{\mathsf{D}}$ and $k \gtrsim \log(1/\delta)/\epsilon^4$ for GoF over $\mathcal{P}_{\mathsf{D}}$ (cf. discussion after Proposition 4.1.3 and Corollary 4.3.4). For the former, recall that the classifier-accuracy test based on $\hat{S}_{1/2}$ achieves the sample complexity

$$n \gtrsim n_{\mathsf{GoF}}(\epsilon, \delta, \mathcal{P}_{\mathsf{D}}) + \frac{k\sqrt{\log(1/\delta)}}{\sqrt{n}\epsilon^2}. \tag{4.3.3}$$

If $n \gtrsim k$ then (4.3.3) is the same as $n \gtrsim n_{\mathsf{GoF}}$; if $m/\log(1/\delta) \lesssim n$ then (4.3.3) is implied by $n \gtrsim n_{\mathsf{GoF}} + \frac{k\log(1/\delta)}{\sqrt{m}\epsilon^2}$, which is optimal within an $O(\log^{1/2}(1/\delta))$ factor (cf. Table 4.2). In our application to GoF we take $m = \infty$, and the missing regime $k \gtrsim \log(1/\delta)/\epsilon^4$ corresponds precisely to $n_{\mathsf{GoF}} \lesssim k$. Summarizing, in the remainder of this section we may assume that $k \wedge (m/\log(1/\delta)) \gtrsim n$.

Let $t = k \wedge (c_0 m/\log(1/\delta))$, where $c_0 > 0$ is a small absolute constant. By the previous paragraph, we assume without loss of generality that $t > n$. For $\ell = \lceil \log_2(t/n) \rceil \geq 1$, define the following $\ell + 2$ subsets of $[k]$:

$$D_0 = \left\{ i : \hat{q}_i^0 \leq \frac{1}{t} \right\}, \quad D_j = \left\{ i : \hat{q}_i^0 \in \left(\frac{2^{j-1}}{t}, \frac{2^j}{t} \right] \right\} \text{ for } j \in [\ell], \quad D_{\ell+1} = \left\{ i : \hat{q}_i^0 > \frac{2^\ell}{t} \right\}.$$

Here $\hat{q}_i^0$ denotes the empirical pmf of $m/2$ held out samples drawn from q (for GoF, one can understand $\hat{q}_i^0 = q_i$ for the distribution q is known). The motivation behind the above choices is the "localization" of each $\hat{q}_i^0$, as shown in the following lemma.

Lemma 4.3.5. *For a small enough universal constant $c_0 > 0$, with probability at least $1 - k\delta$ it holds that for each $i \in [k]$:*

1. *if $\hat{q}_i^0 \in D_0$, then $q_i < 2/t$;*
2. *if $\hat{q}_i^0 \in D_j$ for some $j \in [\ell]$, then $q_i \in (2^{j-2}/t, 2^{j+1}/t]$;*
3. *if $\hat{q}_i^0 \in D_{\ell+1}$, then $q_i > 2^{\ell-1}/t$.*

Lemma 4.3.5 ensures that with high probability, the distribution q restricted to each set D_j is near-uniform. This is similar in spirit to the idea of flattening used in distribution testing [64]. The proof of Lemma 4.3.5 directly follows from the Poisson concentration in Lemma C.1.3 and is thus omitted.

Our main result of this section is the next proposition, which shows that there exist some $j \in \{0, 1, \cdots, \ell + 1\}$ and $\hat{S} \subseteq D_j$ such that $\hat{S}$ is a near-optimal separating set within logarithmic factors.

Proposition 4.3.6. *Suppose $p, q \in \mathcal{P}_{\mathsf{D}}(k)$ with $\mathsf{TV}(p,q) \geq \epsilon$, and X, Y are n iid samples drawn from p, q respectively. There exists a universal constant $c > 0$ such that using the samples X, Y, we can find some $j \in \{0, 1, \cdots, \ell+1\}$ and a set $\hat{S} \subseteq D_j$ which, with probability $1 - \mathcal{O}(k\delta)$, satisfies*

$$\left|\mathrm{sep}(\hat{S})\right| \geq c\left(\frac{\epsilon}{\ell}\right)^2 \left\{ \begin{array}{ll} n/k & \text{if } j = 0 \\ n/\sqrt{kt/2^j} & \text{if } j \in [\ell+1] \end{array} \right\} \qquad \text{and} \quad \tau(\hat{S}) \leq \frac{n2^j}{ct}$$

provided that

$$n\sqrt{1 \wedge \frac{m}{\log(1/\delta)k}} \geq \frac{1}{c} n_{\mathsf{GoF}}(\epsilon/\ell, \delta, \mathcal{P}_{\mathsf{D}}).$$

By Proposition 4.3.6 and Lemma 4.1.2, using the above set $\hat{S}$ leads to the following sample complexity guarantee for the problems GoF and LFHT:

- for GoF under $\mathcal{P}_{\mathsf{D}}$, it succeeds with $n = \Theta(n_{\mathsf{GoF}}(\epsilon/\ell, \delta/k, \mathcal{P}_{\mathsf{D}}))$ observations, which is within a multiplicative $\mathcal{O}(\log^{\Theta(1)}(k))$ factor of the minimax optimal sample complexity in the missing $k \geq \log(1/\delta)/\epsilon^4$ regime;
- for LFHT under $\mathcal{P}_{\mathsf{D}}$ and $m \geq n$, it succeeds with $n = \Theta(n_{\mathsf{GoF}}(\epsilon/\ell, \delta/k, \mathcal{P}_{\mathsf{D}})\sqrt{k\log(k/\delta)/m})$ observations, which is within a multiplicative $\mathcal{O}(\log^{\Theta(1)}(k)\log(k/\delta))$ factor of the minimax optimal sample complexity in the missing $n \leq m \wedge k$.

Therefore, classifier-accuracy tests always lead to near-optimal sample complexities for all GoF, TS, and LFHT problems under both $\mathcal{P}_{\mathsf{Db}}$ and $\mathcal{P}_{\mathsf{D}}$, within polylogarithmic factors in $(k, 1/\delta)$. We leave the removal of extra logarithmic factors for classifier-accuracy tests as an open problem.

4.3.2 The smooth density case

We briefly explain how Corollary 4.3.4 can be used to learn separating sets between distributions in the class $\mathcal{P}_{\mathsf{H}}$ of β-Hölder smooth distributions on $[0,1]^d$. The reduction relies on an approximation result due to Ingster [110, 113], see also [10, Lemma 7.2]. Let P_r be the L^2-projection onto piecewise constant functions on the regular grid on $[0,1]^d$ with r^d cells.

Lemma 4.3.7. *There exist constants c_1, c_2 independent of r such that for any $f \in \mathcal{P}_{\mathsf{H}}(\beta, d, C_{\mathsf{H}})$,*

$$\|P_r f\|_2 \geq c_1\|f\|_2 - c_2 r^{-\beta}.$$

For simplicity write f, g for the Lebesgue densities of $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P}_{\mathsf{H}}$. Suppose $\mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) = \frac{1}{2}\|f - g\|_1 \geq \epsilon$. By Jensen's inequality and Lemma 4.3.7, $\epsilon \lesssim \|P_r(f-g)\|_2$ for $r \asymp \epsilon^{-1/\beta}$. The key observation is that $P_r f$ is essentially the probability mass function of the distribution $\mathbb{P}_{\mathsf{X}}$ when binned on the regular grid with r^d cells. We can now directly apply the results for $\mathcal{P}_{\mathsf{Db}}$ (Corollary 4.3.4) with alphabet size $k \asymp \epsilon^{-d/\beta}$, which combined with Lemma 4.1.2 leads to the sample complexity guarantees in Table 4.1 for the smooth density class $\mathcal{P}_{\mathsf{H}}$ in all three problems GoF, TS and LFHT.

4.3.3 The Gaussian case

Suppose we have two samples X, Y of size n from $\otimes_{j=1}^{\infty}\mathcal{N}(\theta_j^X, 1) =: \mu_{\theta^X}$ and μ_{θ^Y} respectively, where θ^X, θ^Y have Sobolev norm $\|\theta\|_s^2 =: \sum_j \theta_j^2 j^{2s}$ bounded by a constant. In addition, $\mathsf{TV}(\mu_{\theta^X}, \mu_{\theta^Y}) \geq \epsilon > 0$. We use $\hat{\theta}^X$ and $\hat{\theta}^Y$ to denote the empirical mean vector from samples X and Y, respectively.

The separating set is constructed as follows:

$$\hat{S} = \{Z \in \mathbb{R}^{\mathbb{N}} : T(Z) \geq 0\},$$

where $T(Z) = 2\sum_{j=1}^{J}(\hat{\theta}_j^X - \hat{\theta}_j^Y)(Z_j - (\hat{\theta}_j^X + \hat{\theta}_j^Y)/2)$ for some $J \in \mathbb{N}$ to be specified. This is simply a truncated version of the likelihood-ratio test between $\mu_{\hat{\theta}^X}$ and $\mu_{\hat{\theta}^Y}$, where we set all but the first J coordinates of $\hat{\theta}^X$ and $\hat{\theta}^Y$ to zero. The performance of the separating set is summarized in the next proposition.

Proposition 4.3.8. *There exists universal constants c, c' such that when $J = \lfloor c\epsilon^{-1/s} \rfloor$ the inequality*

$$\mathbb{P}\left(\mu_{\theta^X}(\hat{S}) - \mu_{\theta^Y}(\hat{S}) \geq c'\left(\sqrt{n\epsilon^{1/s}} \wedge \frac{1}{\epsilon}\right)\epsilon^2\right) \geq 1 - \delta$$

holds, provided $n \gtrsim \frac{1}{c'} n_{\mathsf{TS}}(\epsilon, \delta, \mathcal{P}_{\mathsf{G}})$.

Applying Proposition 4.3.8 and Lemma 4.1.2 with the trivial bound $\tau(\hat{S}) \leq 1/4$ leads to the sample complexity guarantees in Table 4.1 for the Gaussian sequence model class $\mathcal{P}_{\mathsf{G}}$ in all three problems $\mathsf{GoF}, \mathsf{TS}$ and LFHT.

Chapter 5

Density Estimation Using the Perceptron

This chapter is a reproduction of [75], which is joint work with Tianze Jiang, Yury Polyanskiy and Rui Sun.

5.1 Introduction

A standard step in many machine learning algorithms is to replace an (intractable) optimization over a general function space with an optimization over a large parametric class (most often neural networks). This is done in supervised learning for fitting classifiers, in variational inference [29, 207] for applying ELBO, in variational autoencoders [127] for fitting the decoder, in Generative Adversarial Networks (GANs) [84, 11] for fitting the discriminator, in diffusion models [184, 41] for fitting the score function, and many other settings.

To be specific, let us focus on the example of GANs, which brought about the new era of density estimation in high-dimensional spaces. The problem setting is the following. We are given access to an i.i.d. data $X_1, \dots, X_n \in \mathbb{R}^d$ sampled from an unknown distribution ν and a class of distributions $\mathcal{G}$ on $\mathbb{R}^d$ (the class of available "generators"). The goal of the learner is to find $\arg\min_{\nu' \in \mathcal{G}} D(\nu', \nu)$, where D is some dissimilarity measure ("metric") between probability distributions. In the case of GANs this measure is the Jensen-Shannon divergence $\mathrm{JS}(p, q) \triangleq \mathrm{KL}(p\|\frac{1}{2}p + \frac{1}{2}q) + \mathrm{KL}(q\|\frac{1}{2}p + \frac{1}{2}q)$ where $KL(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} dx$ is the Kullback-Leibler divergence. As any f-divergence, JS has a variational form [see 168, Example 7.5]: $\mathrm{JS}(p, q) = \log 2 + \sup_{h:\mathbb{R}^d \to (0,1)} \mathbb{E}_p[h] + \mathbb{E}_q[\log(1 - h)]$. With this idea in mind, we can now restate the objective of minimizing $\mathrm{JS}(\nu', \nu)$ as a game between a "generator" ν' and a "discriminator" h, i.e., the GAN's estimator is

$$\tilde{\nu} \in \arg\min_{\nu'} \sup_{h:\mathbb{R}^d \to (0,1)} \frac{1}{n} \sum_{i=1}^{n} h(X_i) + \mathbb{E}_{\nu'}[\log(1 - h)], \tag{5.1.1}$$

where we also replaced the expectation over (the unknown) ν with its empirical version $\nu_n =: \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$. Subsequently, the idea was extended to other types of metrics, notably the Wasserstein-GAN [11], which defines

$$\tilde{\nu} \in \arg\min_{\nu' \in \mathcal{G}} \sup_{f \in \mathcal{D}} \left| \mathbb{E}_{Y \sim \nu'} f(Y) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right|, \tag{5.1.2}$$

where the set of discriminators $\mathcal{D}$ is a class of Lipschitz functions (corresponding to the variational characterization of the Wasserstein-1 distance).

The final step to turn (5.1.1) or (5.1.2) into an algorithm is to relax the domain of the inner maximization ("discriminator") to a parametric class of neural network discriminators $\mathcal{D}$. Note that replacing $\sup_{h:\mathbb{R}^d\to(0,1)}$ with $\sup_{h\in\mathcal{D}}$ effectively changes the objective from minimizing the JS divergence to minimizing a "neural-JS", similar to how MINE [23] replaces the true mutual information with a "neural" one. This weakening is quite worrisome for a statistician. While the JS divergence is a strong statistical distance, as it bounds total variation from above and from below [168, Eq. (7.39)], the "neural-JS" is unlikely to possess any such properties.

How does one justify this restriction to a simpler class $\mathcal{D}$? A practitioner would say that while taking $\max_{h\in\mathcal{D}}$ restricts the power of the discriminator, the design of $\mathcal{D}$ is fine-tuned to picking up those features of the distributions that are relevant to the human eye.[1] A theoretician, instead, would appeal to universal approximation results about neural networks to claim that restriction to $\mathcal{D}$ is almost lossless.

The purpose of this paper is to suggest, and prove, a third explanation: the answer is in the *regularity* of ν itself. Indeed, we show that the restriction of discriminators to a very small class $\mathcal{D}$ in (5.1.1) results in almost no loss of minimax statistical guarantees, even if $\mathcal{D}$ is far from being a universal approximator. That is, the minimizing distribution $\tilde{\nu}$ selected with respect to a weak form of the distance enjoys almost minimax optimal guarantees with respect to the strong total variation distance, provided that the true distribution ν is regular enough. Phrased yet another way, even though the "neural" distance is very coarse and imprecise, and hence the minimizer selected with respect to it might be expected to only fool very naive discriminators, in reality it turns out to fool any arbitrarily complex, but bounded discriminator.

Let us proceed to a more formal statement of our results. One may consult Section 5.1.3 for notation. We primarily focus on two classes of distributions on $\mathbb{R}^d$: first, $\mathcal{P}_S(\beta, d, C)$ denotes the set of distributions supported on the d-dimensional unit ball $\mathbb{B}(0,1)$ that have a density with finite L^2 norm and whose $(\beta,2)$-Sobolev norm, defined in (5.1.7), is bounded by C; second, $\mathcal{P}_G(d) = \{\mu * \mathcal{N}(0,1) : \mathrm{supp}(\mu) \subseteq \mathbb{B}(0,1)\}$ is the class of Gaussian mixtures with compactly supported mixing distribution. We remind the reader that the total variation distance has the variational form $\mathsf{TV}(p,q) = \sup_{h:\mathbb{R}^d\to[0,1]} \mathbb{E}_p h - \mathbb{E}_q h$. Our first result concerns the following class of discriminators:

$$\mathcal{D}_1 = \{x \mapsto \mathbb{1}\{x^\top v \geq b\} : v \in \mathbb{R}^d, b \in \mathbb{R}\},$$

the class of affine classifiers, which can be seen as a single layer perceptron with a threshold non-linearity.

Theorem 5.1.1. *For any $\beta > 0$, $d \geq 1$ and $C > 0$, there exists a finite constant C_1 so that*

$$\sup_{\nu\in\mathcal{P}_S(\beta,d,C)} \mathbb{E}\mathsf{TV}(\tilde{\nu},\nu) \leq C_1 n^{-\frac{\beta}{2\beta+d+1}}, \tag{5.1.3}$$

[1] Implying in other words, that whether or not total variation $\mathsf{TV}(\tilde{\nu},\nu)$ is high is irrelevant as long as the generated images look "good enough" to humans.

where the estimator $\tilde{\nu}$ is defined in (5.1.2) with $\mathcal{D} = \mathcal{D}_1$ and $\mathcal{G} = \mathcal{P}_S(\beta, d, C)$. Similarly, for any $d \geq 1$ there exists a finite constant C_2 so that

$$\sup_{\nu \in \mathcal{P}_G(d)} \mathbb{E}\mathsf{TV}(\tilde{\nu}, \nu) \leq C_2 \frac{(\log(n))^{\frac{2d+2}{4}}}{\sqrt{n}},$$

where the estimator $\tilde{\nu}$ is defined in (5.1.2) with $\mathcal{D} = \mathcal{D}_1$ and $\mathcal{G} = \mathcal{P}_G(d)$.

Recall the classical result [107] which shows that the minimax optimal estimation rate in TV over the class $\mathcal{P}_S(\beta, d, C)$ equals $n^{-\beta/(2\beta+d)}$ up to constant factors. Thus, the estimator in (5.1.3) is *almost* optimal, the only difference being that the dimension d is replaced by $d+1$. Similarly, for the Gaussian mixtures we reach the parametric rate up to a polylog factor.[2]

The proof of Theorem 5.1.1 relies on a comparison inequality between total variation and the "perceptron discrepancy", or maximum halfspace distance, which we define as

$$\overline{d_H}(\mu, \nu) =: \sup_{f \in \mathcal{D}_1} \{\mathbb{E}_\mu f - \mathbb{E}_\nu f\}.$$

Note first that $\overline{d_H} \leq \mathsf{TV}$ clearly holds since all functions in the class $\mathcal{D}_1$ are bounded by 1. For the other direction, by proving a generalization of the Gagliardo-Nirenberg-Sobolev inequality we derive the following comparisons.

Theorem 5.1.2. *For any $\beta > 0$, $d \geq 1$ and $C > 0$, there exists a finite constant C_1 so that*

$$\mathsf{TV}(\mu, \nu)^{\frac{2\beta+d+1}{2\beta}} \leq C_1 \overline{d_H}(\mu, \nu) \tag{5.1.4}$$

holds for all $\mu, \nu \in \mathcal{P}_S(\beta, d, C)$. Similarly, for any $d \geq 1$ there exists a finite constant C_2 such that

$$\mathsf{TV}(\mu, \nu) \log\left(3 + \frac{1}{\mathsf{TV}(\mu, \nu)}\right)^{-\frac{d+1}{2}} \leq C_2 \overline{d_H}(\mu, \nu)$$

holds for all $\mu, \nu \in \mathcal{P}_G(d)$.

We remark that we also show (in Proposition 5.3.3) that the exponent $\frac{2\beta+d+1}{2\beta}$ in (5.1.4) is tight, i.e. cannot be improved in general.

With Theorem 5.1.2 in hand the proof of Theorem 5.1.1 is *notably* simple. For example, let us prove (5.1.3) (for full details, see Section 5.4.2). Recall that $X_i \overset{iid}{\sim} \nu$, ν_n is the empirical distribution and $\tilde{\nu} = \arg\min_{\nu' \in \mathcal{P}_S} \overline{d_H}(\nu', \nu_n)$. We then have from the triangle inequality and minimality of $\tilde{\nu}$:

$$\overline{d_H}(\tilde{\nu}, \nu) \leq \overline{d_H}(\tilde{\nu}, \nu_n) + \overline{d_H}(\nu_n, \nu) \leq 2\overline{d_H}(\nu_n, \nu).$$

Thus, from Theorem 5.1.2 we have

$$\mathsf{TV}(\tilde{\nu}, \nu) \leq (2C_1 \overline{d_H}(\nu_n, \nu))^{\frac{2\beta}{2\beta+d+1}}. \tag{5.1.5}$$

[2]For estimation of Gaussian mixtures in total variation the precise value of the minimax optimal polylog factor is at present unknown. However, for the L_2 distance the minimax rate is known, and in the course of our proofs (see (5.3.5)) we show that our estimator only loses a multiplicative factor of $\log(n)^{1/4}$ in loss compared to the optimal L_2-rate $\log(n)^{d/4}/\sqrt{n}$ derived in [125].

Lastly, we recall that $\mathcal{D}_1$ is a class with finite VC-dimension and thus from uniform convergence (Theorem 8.3.23, [201]) we have for some dimension-dependent constant C that

$$\mathbb{E}[\overline{d_H}(\nu_n, \nu)] \le \frac{C}{\sqrt{n}}\,.$$

Thus, applying expectation and Jensen's inequality to (5.1.5) we get

$$\mathbb{E}[\mathsf{TV}(\tilde{\nu}, \nu)] \le C\mathbb{E}[\overline{d_H}(\nu_n, \nu)^{\frac{2\beta}{2\beta+d+1}}] \le C\mathbb{E}[\overline{d_H}(\nu_n, \nu)]^{\frac{2\beta}{2\beta+d+1}} \le Cn^{-\frac{\beta}{2\beta+d+1}}$$

as claimed.

While we believe that Theorem 5.1.1 provides theoretical proof for the efficacy of simple discriminators, it has several serious theoretical and practical deficiencies that we now address. First, the rate for the class $\mathcal{P}_S$ is not minimax optimal. In this regard, we show that by replacing the perceptron class $\mathcal{D}_1$ with a generalized perceptron

$$\mathcal{D}_\gamma = \{x \mapsto |x^\top v - b|^{\frac{\gamma-1}{2}} : v \in \mathbb{R}^d, b \in \mathbb{R}\}\,, \qquad \gamma \in (0, 2)$$

and taking an *average* over $\mathcal{D}_\gamma$ instead of a supremum, we are able to achieve a total variation rate of $n^{-\beta/(2\beta+d+\gamma)}$, thus coming arbitarily close to minimax optimality as $\gamma \to 0$. See Theorem 5.4.1 for details.[3]

Second, from the implementation point of view, the density estimation algorithm behind Theorem 5.1.1 is completely impractical. Indeed, finding the halfspace with maximal separation between even two empirical measures is a nonconvex, non-differentiable problem and takes super-poly time in the dimension d assuming $\mathsf{P} \neq \mathsf{NP}$ [89], and $\omega(d^{\omega(\epsilon^{-1})})$ time for ϵ-optimal agnostic learning between two densities assuming either SIVP or gapSVP [197].

Even if we disregard the computational complexity, it is unclear how to find the exact minimizer $\tilde{\nu}$ of $\arg\min_{\nu'} \overline{d_H}(\nu', \nu_n)$. This concern is alleviated by the fact that any $\tilde{\nu}$ satisfying $\overline{d_H}(\tilde{\nu}, \nu_n) = O(\sqrt{d/n})$ will work withouth degrading our performance guarantee, and thus only an approximate minimizer is needed. Taking this one step further, our proof proceeds by replacing the perceptron discrepancy $\overline{d_H}$ (defined with respect to the best perceptron) with an average version d_H defined in (5.2.2), for which the comparison in Theorem 5.1.2 still holds. Therefore, one does not even need to find an approximately optimal half-space, as random half-spaces provide sufficient discriminatory power.

Somewhat unexpectedly, we discover that the average perceptron discrepancy d_H exactly equals Székely and Rizzo's energy distance $\mathcal{E}_1$ (Definition 1, [190]), defined as

$$\mathcal{E}_1^2(\mu, \nu) \triangleq \mathbb{E}\left[2\|X - Y\| - \|X - X'\| - \|Y - Y'\|\right], \qquad (X, X', Y, Y') \sim \mu^{\otimes 2} \otimes \nu^{\otimes 2}\,, \quad (5.1.6)$$

where $\|\cdot\|$ is the usual Euclidean norm on $\mathbb{R}^d$. Thus, our Theorem 5.4.1 (with $\gamma = 1$) shows that minimizing $\min_{\nu'} \mathcal{E}_1(\nu', \nu_n)$ gives a density estimator with rates over $\mathcal{P}_S$ and $\mathcal{P}_G$ as given in Theorem 5.1.1.

From the algorithmic point of view our message is the following. If one has access to a parametric family of generators sampling from ν_θ for parameters $\theta \in \mathbb{R}^p$, and if one can compute ∇_θ of the generator forward pass, e.g., via pushforward of a reference distribution under a smooth transport map or neural network-based models [202, 147], then one can fit θ to the empirical sample ν_n by running stochastic gradient descent steps:

[3] More precisely, within a polylog in n.

- sample m samples from ν_θ and form the empirical distribution ν'_m,
- compute the loss $\mathcal{E}_1(\nu'_m, \nu_n)$ and backpropagate the gradient with respect to θ,
- update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{E}_1(\nu'_m, \nu_n)$ for some step size η.

Note that the computation of $\mathcal{E}_1(\nu'_m, \nu_n)$ according to (5.1.6) requires $O(n^2 + m^2)$ steps and is friendly to gradient evaluations.

5.1.1 Contributions

To summarize, our main contributions are as follows. We show that β-smooth distributions, Gaussian mixtures and discrete distributions that are far apart in total variation distance must possess a halfspace on which their mass is substantially different (Theorems 5.1.2, 5.3.2 and 5.3.5).

We apply the separation results to density estimation problems, showing that an ERM density estimator nearly attains the minimax optimal density estimation rate with respect to TV over the aforementioned distribution classes (Theorems 5.1.1, 5.4.1 and 5.4.4).

In Section 5.2 we show that the average halfspace separation distance d_H is equal up to constant to the energy distance $\mathcal{E}_1$ (Proposition 5.2.2), which has many equivalent expressions: as a weighted L^2-distance between characteristic functions (Proposition 5.2.3), as the sliced Cramér-2 distance (Proposition 5.2.5), as an IPM/MMD/energy distance (Section 5.2.3), and as the L^2-norm of the Riesz potential (Proposition 5.2.6).

We generalize the average halfspace distance d_H to include an exponent $\gamma \in (0, 2)$, corresponding to the generalized energy distance $\mathcal{E}_\gamma$. Consequently, we discover that if instead of thresholded linear features $\mathbb{1}\{v^\top x > b\}$ we use the non-linearity $|v^\top x - b|^\gamma$, smooth distributions and Gaussian mixtures can be separated even better (Theorem 5.3.2). Combined with the fact that $\mathcal{E}_\gamma$, similarly to d_H, decays between population and sample measures at the parametric rate (Lemma 5.2.4), the ERM for $\mathcal{E}_\gamma$ reduces the slack in the density estimation rate, almost achieving minimax optimality. This result, combined with its strong approximation properties, supports its use in modern generative models (e.g. [95, 84, 176, 172]).

Finally, Proposition 5.5.1 shows that recent work applying $\overline{d_H}$ for two-sample testing is sub-optimal over the class of smooth distributions in the minimax sense.

5.1.2 Related Work

In the statistics literature, an estimator of the form (5.1.2) appears in the famous work of [206]. Instead of indicators of halfspaces, they consider the class of discriminators

$$\mathcal{Y}_\epsilon =: \{\mathbb{1}\{\mathrm{d}\nu_i/\mathrm{d}\nu_j \geq 1\} : 1 \leq i, j \leq N(\epsilon, \mathcal{G})\},$$

where $\nu_1, \ldots, \nu_{N(\epsilon,\mathcal{G})}$ forms a minimal ϵ-TV covering of the class $\mathcal{G}$ and $N(\epsilon, \mathcal{G})$ is the so-called covering number. Writing $d_Y(\mu, \mu') = \sup_{f \in \mathcal{Y}}(\mathbb{E}_\mu f - \mathbb{E}_{\mu'} f)$, it is not hard to prove that $|\mathsf{TV} - d_Y| = O(\epsilon)$ on $\mathcal{G} \times \mathcal{G}$ and that $\mathbb{E} d_Y(\nu, \nu_n) \lesssim \sqrt{\log N(\epsilon_n, \mathcal{G})/n}$ by a union bound coupled with a binomial tail inequality. From here $\mathbb{E}\mathsf{TV}(\tilde{\nu}, \nu) \lesssim \sqrt{\log N(\epsilon, \mathcal{G})/n} + \epsilon$ follows by the

triangle inequality (here $\tilde{\nu}$ is defined as in (5.1.2) with $\mathcal{D} = \mathcal{Y}$). Note that in contrast to our perceptron discrepancy $\overline{d_H}$, Yatracos' estimator attains the optimal rate on $\mathcal{G} = \mathcal{P}_S$, corresponding to the choice $\epsilon = \epsilon(n) \asymp n^{-\beta/(2\beta+d)}$.

The paper by [160] derives minimax density estimation guarantees for a class of diffusion based estimators. Their result is similar to ours in that it obtains rigorous (near-)optimal guarantees for a method that is a realistic model of what is currently done in practice. However, their analysis relies crucially on the universal approximation property of neural networks, as they must show that the true score function can be approximated to high precision. This stands in contrast to the present paper: the discriminator class of halfspace indicators that we study certainly does not have such a universal approximation property. However, our analysis does require the generator network to be able to fit the data sufficiently well.

Several other related works such as [181, 141] study the problem of minimax density estimation over classical smoothness classes with respect to Integral Probability Metrics (IPMs) $d_{\mathcal{D}}(\mathbb{P}, \mathbb{Q}) =: \sup_{f \in \mathcal{D}} |\mathbb{E}_{\mathbb{P}} f - \mathbb{E}_{\mathbb{Q}} f|$. In particular, these works seek estimators $\tilde{\nu}$ such that $d_{\mathcal{D}}(\tilde{\nu}, \nu)$ is small for some discriminator $\mathcal{D}$. Note some crucial differences to our work: first, we evaluate performance with respect to total variation in Theorem 5.1.1 which bears more interest both theoretically and empirically; second, we restrict our attention to estimators $\tilde{\nu}$ attained by ERM which is more commonly used in practice.

A paper closer in spirit to ours is [12] whose authors study comparison inequalities between the Wasserstein distance W_1 and the IPM d_{relu} defined by the discriminator class $\mathcal{D} = \{x \mapsto \mathrm{Relu}(x^\top v + b) : b, \|v\| \leq 1\}$. They show [12, Theorem 3.1] that $\sqrt{\kappa/d} W_1 \lesssim d_{relu} \lesssim W_1$ for Gaussian distributions with mean in the unit ball, where κ is an upper bound on their condition numbers and d is the dimension. They obtain results for other distribution classes (Gaussian mixtures, exponential families), but for each of these they use a different class of discriminators that is adapted to the problem. In contrast, we mainly focus the discriminator class $\mathcal{D}_1 = \{x \mapsto \mathbb{1}\{x^\top v \geq b\} : \|v\| \leq 1, b \in \mathbb{R}\}$ and are able to derive novel comparisons to TV for smooth distributions, Gaussian mixtures and discrete distributions. In addition, we prove the (near)-optimality of our results (for smooth densities) and also derive nonparametric estimation rates for the corresponding GAN density estimators.

Independent of this work, recent results by [161] investigate the halfspace separability of distributions for the setting of two-sample testing. However, their focus was on the asymptotic power of the test as the number of samples grows to infinity. Our lower bound construction presented in Appendix D.4 proves that their proposed test is sub-optimal in the minimax setting. See Section 5.5 for a more detailed discussion.

5.1.3 Notation

The symbols $O, o, \Theta, \Omega, \omega$ follow the conventional "big-O" notation, and $\tilde{O}, \tilde{o}$ hide polylogarithmic factors. We use $\lesssim$, $\gtrsim$ and $\asymp$ throughout our calculations to hide multiplicative constants that are irrelevant (depending on the context). Given a vector $x \in \mathbb{R}^d$, we write $\|x\|$ for its Euclidean norm and $\langle x, y \rangle =: x^\top y$ for the Euclidean inner product of $x, y \in \mathbb{R}^d$. The Gamma function is denoted by Γ. We write $\mathbb{B}(x, r) =: \{y \in \mathbb{R}^d : \|x - y\| \leq r\}$, $\mathbb{S}^{d-1} =: \{x \in \mathbb{R}^d : \|x\| = 1\}$ and σ for the unnormalized surface measure on $\mathbb{S}^{d-1}$. The surface

area of a unit $(d-1)$-sphere is also written as $\sigma(\mathbb{S}^{d-1}) = 2\pi^{d/2}/\Gamma\left(\frac{d}{2}\right)$. In particular, if X is a random vector uniformly distributioned on $\mathbb{S}^{d-1}$ then for any h we have

$$\mathbb{E}[h(X)] = \frac{1}{\sigma(\mathbb{S}^{d-1})} \int_{\mathbb{R}^d} h(y) \mathrm{d}\sigma(y) \,.$$

The convolution between functions/measures is denoted by $*$. We write $L^p(\mathbb{R}^d)$ for the space of (equivalence classes of) functions $\mathbb{R}^d \to \mathbb{C}$ that satisfy $\|f\|_p =: \left(\int_{\mathbb{R}^d} |f(x)|^p \mathrm{d}x\right)^{1/p} < \infty$. The space of all probability distributions on $\mathbb{R}^d$ is denoted as $\mathcal{P}(\mathbb{R}^d)$. For a signed measure ν we write $\mathrm{supp}(\nu)$ for its support and $M_r(\nu) =: \int \|x\|^r \mathrm{d}|\nu|(x)$ for its r'th absolute moment. Given $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathbb{R}^d)$ we write $\mathsf{TV}(\mathbb{P}, \mathbb{Q}) =: \sup_{A \subseteq \mathbb{R}^d} [\mathbb{P}(A) - \mathbb{Q}(A)]$ for the total variation distance, where the supremum is over all measurable sets.

Given a function $f \in L^1(\mathbb{R}^d)$, define its Fourier transform as

$$\widehat{f}(\omega) =: \mathcal{F}[f](\omega) =: \int_{\mathbb{R}^d} e^{-i\langle x, \omega \rangle} f(x) \mathrm{d}x.$$

Given a finite signed measure ν on $\mathbb{R}^d$, define its Fourier transform as $\mathcal{F}[\nu](\omega) =: \int_{\mathbb{R}^d} e^{-i\langle \omega, x \rangle} \mathrm{d}\nu(x)$. We extend the Fourier transform to $L^2(\mathbb{R}^d)$ and tempered distributions in the standard manner. Given $f \in L^2(\mathbb{R}^d)$ and $\beta > 0$, define its homogenous Sobolev seminorm of order $(\beta, 2)$ as

$$\|f\|_{\beta,2}^2 =: \int_{\mathbb{R}^d} \|\omega\|^{2\beta} |\widehat{f}(\omega)|^2 \mathrm{d}\omega. \tag{5.1.7}$$

Further, we define two specific classes of functions of interest as follows: $\mathcal{P}_S(\beta, d, C)$ is a set of smooth densities while $\mathcal{P}_G(d)$ is a set of all Gaussian mixtures with support in the unit ball, formally

$$\begin{aligned} \mathcal{P}_S(\beta, d, C) &=: \{\mu \in \mathcal{P}(\mathbb{R}^d) : \mathrm{supp}(\mu) \subseteq \mathbb{B}(0,1), \mu \text{ has density } p \text{ with } \|p\|_{\beta,2} \leq C\}, \\ \mathcal{P}_G(d) &=: \{\nu * \mathcal{N}(0, I_d) : \nu \in \mathcal{P}(\mathbb{R}^d), \mathrm{supp}(\nu) \subseteq \mathbb{B}(0,1)\}. \end{aligned}$$

Assumption 3. *Throughout the paper we assume that C in the definition of $\mathcal{P}_S(\beta, d, C)$ is large enough relative to β and d, such that $\mathcal{P}_S(\beta, d, C/2)$ is non-empty.*

5.1.4 Structure

The structure of the paper is as follows. In Section 5.2 we introduce the generalized energy distance, the main object of our study. We show how it relates to the perceptron discrepancy $\overline{d_H}$ and its relaxation d_H; we record equivalent formulations of the generalized energy distance, one of which is a novel "sliced-distance" form. In Section 5.3, we present our main technical results on comparison inequalities between total variation and the energy distance. In Section 5.4 we analyse the density estimator that minimizes the empirical energy distance, and prove Theorem 5.1.1 and Theorem 5.1.2 in Section 5.4.2. In Section 5.5 we show that the use of $\overline{d_H}$ for two sample testing results in suboptimal performance. We conclude in Section 5.6. All omitted proofs and auxiliary results are deferred to the Appendix.

5.2 The Generalized Energy Distance

Given two probability distributions μ, ν on $\mathbb{R}^d$ with finite γ'th moment, the generalized energy distance of order $\gamma \in (0, 2)$ between them is defined as

$$\mathcal{E}_\gamma(\mu, \nu) = \mathbb{E}\Big[2\|X - Y\|^\gamma - \|X - X'\|^\gamma - \|Y - Y'\|^\gamma\Big], \qquad \text{where } (X, X', Y, Y') \sim \mu^{\otimes 2} \otimes \nu^{\otimes 2}. \tag{5.2.1}$$

As we alluded to in the introduction, the proof of Theorems 5.1.1 and 5.1.2 becomes possible once we relax the supremum in the definition of $\overline{d_H}$ to an *unnormalized* average over halfspaces. In Section 5.2.1 we discuss this relaxation in more detail and identify a connection to the energy distance $\mathcal{E}_1$ defined above in (5.2.1). Motivated by this, we study the (generalized) energy distance and give multiple equivalent characterizations of it from Section 5.2.1 to Section 5.2.5.

5.2.1 From Perceptron Discrepancy to Energy Distance

Our first goal is to connect the study of $\overline{d_H}$ to the study of $\mathcal{E}_\gamma$ with $\gamma = 1$. To achieve this, we introduce an intermediary, the "average" perceptron discrepancy d_H. Given two probability distributions μ, ν on $\mathbb{R}^d$, we define

$$d_H(\mu, \nu) =: \sqrt{\int_{v \in \mathbb{S}^{d-1}} \int_{b \in \mathbb{R}} \left(\int_{\langle v, x \rangle \geq b} \mathrm{d}\mu(x) - \mathrm{d}\nu(x) \right)^2 \mathrm{d}b \mathrm{d}\sigma(v)}, \tag{5.2.2}$$

where σ denotes the surface area measure.

If the two distributions μ, ν are supported on a compact set, then the overall definition can indeed be regarded as a 'mean squared' version of perceptron discrepancy, because the integrals over b and v only range over bounded sets. However, in general, the integral over b in the definition of d_H is not normalizable and that is why we put "average" in quotes. Nevertheless, we have the following comparisons between d_H and $\overline{d_H}$.

Proposition 5.2.1. *For any $\beta > 0$, $d \geq 1$, $C > 0$, and for all $\mu, \nu \in \mathcal{P}_S(\beta, d, C)$, we have*

$$\sqrt{\frac{\Gamma(d/2)}{4\pi^{d/2}}} d_H(\mu, \nu) \leq \overline{d_H}(\mu, \nu). \tag{5.2.3}$$

Moreover, for all $d \geq 1$, there exists a finite constant C_1 such that for all $\mu, \nu \in \mathcal{P}_G(d)$,

$$\frac{d_H(\mu, \nu)}{\log(3 + 1/d_H(\mu, \nu))^{1/4}} \leq C_1 \overline{d_H}(\mu, \nu).$$

Proof. The proof of (5.2.3) is immediate after noting that all distributions in $\mathcal{P}_S(\beta, d, C)$ are supported on the d-dimensional unit ball and that $\int_{v \in \mathbb{S}^{d-1}} \int_{-1}^{1} \mathrm{d}b \mathrm{d}\sigma(v) = 4\pi^{d/2}/\Gamma(d/2)$. Thus, we focus on the Gaussian mixture case. Write $\mu - \nu = \tau * \phi$ where ϕ denotes the density of the standard Gaussian $\mathcal{N}(0, I_d)$ and $\tau \in \mathcal{P}(\mathbb{R}^d)$ is the difference of the two implicit mixing

measures. For any $R > 0$, we have

$$\begin{aligned}\overline{d_H}(\mu,\nu) &\geq \sup_{v\in\mathbb{S}^{d-1},|b|\leq R}\int_{\langle x,v\rangle\geq b}(\tau*\phi)(x)\mathrm{d}x \\ &\geq \sqrt{\frac{1}{2R\,\mathrm{vol}_{d-1}(\mathbb{S}^{d-1})}\int_{\mathbb{S}^{d-1}}\int_{|b|\leq R}\left(\int_{\langle x,v\rangle\geq b}(\tau*\phi)(x)\mathrm{d}x\right)^2\mathrm{d}b\mathrm{d}\sigma(v)}.\end{aligned}$$

Now, since τ is supported on a subset of $\mathbb{B}(0,1)$ by definition of the class $\mathcal{P}_G(d)$, for any $v\in\mathbb{S}^{d-1}$ and $R\geq 2$ we have the bound

$$\begin{aligned}\int_{|b|>R}\left(\int_{\langle x,v\rangle\geq b}\int_{\mathbb{R}^d}\phi(x-y)\mathrm{d}\tau(y)\mathrm{d}x\right)^2\mathrm{d}b &\leq \int_{|b|>R}\left(\int_{\langle x,v\rangle\geq|b|}\exp(-(\|x\|-1)^2/2)\mathrm{d}x\right)^2\mathrm{d}b \\ &\leq \int_{|b|>R}\left(\int_{\|x\|\geq|b|}\exp(-\|x\|^2/8)\mathrm{d}x\right)^2\mathrm{d}b \\ &\lesssim \exp(-\Omega(R^2)),\end{aligned}$$

where we implicitly used that $\int\mathrm{d}\tau = 0$ as τ is the difference of two probability distributions. Choosing $R\asymp\sqrt{\log(3+1/d_H(\mu,\nu))}$ concludes the proof. □

Proposition 5.2.1 implies that to obtain a comparison between TV and $\overline{d_H}$, specifically for lower bounding $\overline{d_H}$, it suffices to consider the relaxation d_H instead. The next observation we make is that d_H is in fact equal, up to constant, to the energy distance.

Proposition 5.2.2. *Let μ,ν be probability distributions on $\mathbb{R}^d$ with finite mean. Then*

$$d_H(\mu,\nu) = \frac{\pi^{(d-1)/4}}{\sqrt{\Gamma(\frac{d+1}{2})}}\mathcal{E}_1(\mu,\nu).$$

Proof. This is a direct implication of (5.2.4) and (5.2.7). We defer the proof to the more general Proposition 5.2.5. □

As will be clear from the rest of the paper, it does pay off to study $\mathcal{E}_\gamma$ for general γ, even though so far we only justified its relevance to the results stated in the introduction for the case of $\gamma = 1$. With this in mind, we proceed to study various properties of the *generalized* energy distances $\{\mathcal{E}_\gamma\}_{\gamma\in(0,2)}$.

5.2.2 The Fourier Form

The formulation of the generalized energy distance that we rely on most heavily in our proofs is the following.

Proposition 5.2.3 ([190, Proposition 2])**.** *Let $\gamma\in(0,2)$ and let μ,ν be probability distributions on $\mathbb{R}^d$ with finite γ'th moment. Then,*

$$\mathcal{E}_\gamma^2(\mu,\nu) = F_\gamma(d)\int_{\mathbb{R}^d}\frac{|\widehat{\mu}(\omega)-\widehat{\nu}(\omega)|^2}{\|\omega\|^{d+\gamma}}\mathrm{d}\omega, \tag{5.2.4}$$

where we define $F_\gamma(d) = \frac{\gamma 2^{\gamma-1}\Gamma(\frac{d+\gamma}{2})}{\pi^{d/2}\Gamma(1-\frac{\gamma}{2})}$.

Remark 18. *Note that* $F_\gamma(d) = \Theta\left(\gamma(2-\gamma)\Gamma\left(\frac{d+\gamma}{2}\right)\pi^{-d/2}\right)$ *up to a universal constant.*

This shows that the generalized energy distance is a weighted L^2 distance in Fourier space. The fact that $\mathcal{E}_\gamma$ is a valid metric on probability distributions with finite γ'th moment is a simple consequence of Proposition 5.2.3.

5.2.3 The MMD and IPM Forms

Another interpretation of the generalized energy distance is through the theory of *Maximum Mean Discrepancy* (MMD). Given a set $\mathcal{X}$ and a positive semidefinite kernel $k : \mathcal{X}^2 \to \mathbb{R}$, there is a unique reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ consisting of the closure of the linear span of $\{k(x,\cdot), x \in \mathcal{X}\}$ with respect to the inner product $\langle k(x,\cdot), k(y,\cdot)\rangle_{\mathcal{H}_k} = k(x,y)$.

For a probability distribution μ on $\mathcal{X}$, define its kernel embedding as $\theta_\mu = \int_{\mathbb{R}^d} k(x,\cdot)\mathrm{d}\mu(x)$. As shown in [151, Lemma 3.1], the kernel embedding θ_μ exists and belongs to the RKHS $\mathcal{H}_k$ if $\mathbb{E}[\sqrt{k(X,X')}] < \infty$ for $(X,X') \sim \mu^{\otimes 2}$ — as is the case for our kernel defined later in Equation (5.2.5). Then, given two probability distributions μ and ν, the MMD measures their distance in the RKHS by

$$\mathrm{MMD}_k(\mu,\nu) =: \|\theta_\mu - \theta_\nu\|_{\mathcal{H}_k}.$$

We refer the reader to [179, 151] for more details on the underlying theory. MMD has a closed form thanks to the reproducing property:

$$\mathrm{MMD}_k^2(P,Q) = \mathbb{E}\Big[k(X,X') + k(Y,Y') - 2k(X,Y)\Big],$$

where $(X,X',Y,Y') \sim \mu^2 \otimes \nu^2$. Moreover, it also follows that MMD is an *Integral Probability Metric (IPM)* where the supremum is over the unit ball of the RKHS $\mathcal{H}_K$:

$$\mathrm{MMD}_k(\mu,\nu) = \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \le 1} \mathbb{E}[f(X) - f(Y)].$$

In our case, we can define the kernel

$$k_\gamma(x,y) = \|x\|^\gamma + \|y\|^\gamma - \|x-y\|^\gamma \tag{5.2.5}$$

for $\gamma \in (0,2)$, which in one dimension corresponds to the covariance operator of fractional Brownian motion. For a proof of the nontrivial fact that k_γ above is positive definite see for example [180]. With the choice of k_γ it follows trivially from its definition that the generalized energy distance $\mathcal{E}_\gamma$ is equal to the MMD with kernel k_γ, i.e.

$$\mathcal{E}_\gamma(\mu,\nu) = \mathrm{MMD}_{k_\gamma}(\mu,\nu)$$

for all distributions μ, ν with finite γ'th moment. It is noteworthy that while $\overline{d_H}$ is by definition an IPM, so is its averaged version d_H.

A straightforward consequence of the above characterization is the fact that $\mathcal{E}_\gamma$ decays at the parametric rate between empirical and population measures. This is not terribly surprising as analogous results hold for arbitrary MMDs with bounded kernel, see for example [85, Theorem 7]. Recall that $M_t(\nu)$ denotes the t'th absolute moment of the measure ν.

Lemma 5.2.4. *Let ν be a probability distribution on $\mathbb{R}^d$ and let $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ for an i.i.d. sample $X_1, \dots, X_n$ from ν. Then, for any $\gamma \in (0,2)$,*

$$\mathbb{E}\mathcal{E}_\gamma^2(\nu, \nu_n) \le \frac{2M_\gamma(\nu)}{n}.$$

For a high-probability bound when ν is compactly supported, see Lemma 5.4.5.

Proof. Let $\tilde{X}_1, \dots, \tilde{X}_n$ be an additional i.i.d. sample from ν, and write $\tilde{\nu}_n$ for the corresponding empirical measure. Using the definition of $\mathcal{E}_\gamma$ in (5.2.1), we can compute

$$\begin{aligned}\mathbb{E}\mathcal{E}_\gamma^2(\nu_n, \tilde{\nu}_n) &= \mathbb{E}\Big[\frac{2}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|X_i - \tilde{X}_j\|^\gamma - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|\tilde{X}_i - \tilde{X}_j\|^\gamma \\ &\qquad - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|X_i - X_j\|^\gamma\Big] \\ &= \frac{2}{n}\mathbb{E}\|X_1 - X_2\|^\gamma.\end{aligned}$$

The conclusion then follows from taking the expectation of the expression

$$\mathbb{E}\Big[\mathcal{E}_\gamma^2(\tilde{\nu}_n, \nu_n)\Big|\nu_n\Big] = \mathcal{E}_\gamma^2(\nu, \nu_n) + \frac{1}{n}\mathbb{E}\|X_1 - X_2\|^\gamma$$

and the inequality $|x+y|^\gamma \le 2^{\max\{0,\gamma-1\}}(|x|^\gamma + |y|^\gamma)$ for all $x, y \in \mathbb{R}$. □

5.2.4 The Sliced Form

Another equivalent characterization of the generalized energy distance is in the form of a *sliced* distance. Sliced distances are calculated by first choosing a random direction on the unit sphere, and then computing a one-dimensional distance in the chosen direction between the projections of the two input distributions. For $\gamma \in (0,2)$ define the function

$$\psi_\gamma(x) = \begin{cases} |x|^{(\gamma-1)/2} & \text{for } \gamma \ne 1 \\ \mathbb{1}\{x \ge 0\} & \text{otherwise.} \end{cases} \tag{5.2.6}$$

The following result, to the best of our knowledge, has not appeared in prior literature except for the case of $\gamma = 1$.

Proposition 5.2.5. *Let $\gamma \in (0,2)$ and let μ, ν be probability distributions on $\mathbb{R}^d$ with finite γ'th moment. Then for $(X,Y) \sim \mu \otimes \nu$ we have*

$$\mathcal{E}_\gamma^2(\mu,\nu) = \frac{1}{S_\gamma}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}}\Big[\mathbb{E}\psi_\gamma(\langle X, v\rangle - b) - \mathbb{E}\psi_\gamma(\langle Y, v\rangle - b)\Big]^2 \mathrm{d}b\mathrm{d}\sigma(v), \tag{5.2.7}$$

where $S_\gamma = \frac{\pi^{\frac{d}{2}+1}\Gamma(1-\frac{\gamma}{2})}{\gamma 2^{\gamma-1}\Gamma(\frac{d+\gamma}{2})\cos^2(\frac{\pi(\gamma-1)}{4})\Gamma(\frac{1-\gamma}{2})^2}$ *when* $\gamma \ne 1$ *and* $S_1 = \frac{\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d+1}{2})}$.

The proof of Proposition 5.2.5 hinges on computing the Fourier transform of the function ψ_γ, which can be interpreted as a tempered distribution. We point out a special property of the integral on the right hand side of (5.2.7). After expanding the square, one finds that the individual terms in the sum are not absolutely integrable for $\gamma \neq 1$. However, due to cancellations within the squared quantity, the integral is finite.

As claimed, using the language of [152], Proposition 5.2.5 allows us to interpret $\mathcal{E}_\gamma$ as a *sliced* probability divergence. Given $v \in \mathbb{S}^{d-1}$, write $\theta_v = \langle v, \cdot \rangle$ and $\theta_v \# \nu = \nu \circ \theta_v$ for the pushforward of ν under θ_v. We have

$$S_\gamma(d)\mathcal{E}_\gamma^2(\mu,\nu) = S_\gamma(1)\int_{\mathbb{S}^{d-1}} \mathcal{E}_\gamma^2(\theta_v\#\mu, \theta_v\#\nu)\mathrm{d}\sigma(v).$$

We may also observe that the energy distance $\mathcal{E}_1$ is equal to the sliced Cramér-2 distance up to constant, which has been studied recently by both theoretical and empirical works [130, 131].[4]

5.2.5 The Riesz Potential Form

The generalized energy distance can also be linked to the Riesz potential [138, Chapter 1.1], which is the inverse of the fractional Laplace operator. Given $0 < s < d$, the Riesz potential $I_s f$ of a compactly supported signed measure f on $\mathbb{R}^d$ is defined (in a weak sense) by

$$I_s f = f * K_s,$$

where $K_s(x) = c_s^{-1}\|x\|^{s-d}$ and $c_s = \pi^{d/2}2^s\Gamma(s/2)\Gamma((d-s)/2)^{-1}$. The Fourier transform of the Riesz kernel is given by $\widehat{K_s}(\omega) = \|\omega\|^{-s}$, interpreted as a tempered distribution. The following proposition is derived by setting $s = \frac{d+\gamma}{2}$ and using the Fourier form (Proposition 5.2.3) of the energy distance.

Proposition 5.2.6. *Let $\gamma \in (0, \min\{d, 2\})$ and let μ, ν be compactly supported probability distributions on $\mathbb{R}^d$. Then*

$$\mathcal{E}_\gamma(\mu,\nu) = (2\pi)^{d/2}\sqrt{F_\gamma(d)}\|I_{\frac{d+\gamma}{2}}(\mu-\nu)\|_2. \tag{5.2.8}$$

5.3 Main Comparison: TV Versus Energy

After considering the connection of the perceptron discrepancy $\overline{d_H}$ to the energy distance in Section 5.2, we turn to some of our main technical results, which provide novel quantitative comparisons between $\{\mathcal{E}_\gamma\}_{\gamma\in(0,2)}$ and the total variation distance. In Section 5.3.1 we show that the generalized energy distance is upper bounded by total variation for compactly supported distributions. In Section 5.3.2 we derive lower bounds on the generalized energy distance in terms of the total variation distance over the two distribution classes that we have introduced, namely smooth distributions and Gaussian mixtures. Finally, in Section 5.3.3 we turn to the case of discrete distributions, which requires alternative techniques.

[4]The Cramér-p distance is simply the L^p distance between cumulative distribution functions.

5.3.1 Upper Bound — Compactly Supported Distributions

Note that we (obviously) always have $\overline{d_H}(\mu,\nu) \le \mathsf{TV}(\mu,\nu)$ for arbitrary probability measures μ and ν. Moreover, for distributions supported on a unit ball we also have $d_H(\mu,\nu) \lesssim \overline{d_H}(\mu,\nu)$. Therefore, by the identification of d_H and $\mathcal{E}_1$ (Proposition 5.2.2), we can see that for distributions with bounded support, we always have $\mathcal{E}_1(\mu,\nu) \lesssim \mathsf{TV}(\mu,\nu)$. The next result generalizes this estimate for all $\mathcal{E}_\gamma$, not just $\gamma = 1$.

Proposition 5.3.1. *For any dimension $d \ge 1$ and $\gamma \in (0, \min\{d,2\})$ there exists a finite constant c such that for any two probability distributions μ,ν supported on the unit ball we have*

$$\mathcal{E}_\gamma(\mu,\nu) \le c\mathsf{TV}(\mu,\nu)\,.$$

Proof. Assume first that both μ,ν are absolutely continuous, and let $f(x) = \frac{\mathrm{d}\mu}{\mathrm{d}x} - \frac{\mathrm{d}\nu}{\mathrm{d}x}$ and $\epsilon = \mathsf{TV}(\mu,\nu)$. By Equation (5.2.8), it suffices to upper bound $\|I_s f\|_2$ for $s = (d+\gamma)/2$. First we decompose $\|I_s f\|_2$ as

$$\|I_s f\|_2 \le \|I_s f \mathbb{1}_{\mathbb{B}(0,2)^c}\|_2 + \|I_s f \mathbb{1}_{\mathbb{B}(0,2)}\|_2$$

by the triangle inequality. Let $f^+(x) = \max\{f(x),0\}$ and $f^-(x) = \max\{-f(x),0\}$ so that $\int f^+(x)\mathrm{d}x = \int f^-(x)\mathrm{d}x = \epsilon$. Since $\mathrm{supp}(f) \subseteq \mathbb{B}(0,1)$, for all $\|x\| > 2$ we have

$$\begin{aligned} c_s I_s f(x) &= \int \frac{f^+(y) - f^-(y)}{\|x-y\|^{d-s}}\mathrm{d}y \\ &\le \frac{\int f^+(y)\mathrm{d}y}{(\|x\|-1)^{d-s}} - \frac{\int f^-(y)\mathrm{d}y}{(\|x\|+1)^{d-s}} \\ &= \epsilon\left(\frac{1}{(\|x\|-1)^{d-s}} - \frac{1}{(\|x\|+1)^{d-s}}\right) \\ &\le \epsilon\frac{2(d-s)}{(\|x\|-1)^{d-s+1}} \end{aligned}$$

where the last line follows from the convexity of the function $u \mapsto u^{s-d}$ for $u > 0$. Thus, we can upper bound $\|I_s f\, \mathbb{1}_{\mathbb{B}(0,2)^c}\|_2$ by

$$\begin{aligned} \|I_s f\, \mathbb{1}_{\mathbb{B}(0,2)^c}\|_2 &\le \frac{\epsilon}{c_s}\left(\int_{\mathbb{B}(0,2)^c} \frac{4(d-s)^2}{(\|x\|-1)^{2d-2s+2}}\mathrm{d}x\right)^{1/2} \\ &= \frac{2(d-s)\sqrt{\sigma(\mathbb{S}^{d-1})}\epsilon}{c_s}\left(\int_2^\infty \frac{u^{d-1}}{(u-1)^{2d-2s+2}}\mathrm{d}u\right)^{1/2} \lesssim \epsilon \end{aligned}$$

where we discard a finite constant depending only on d and γ.

Next we need to estimate $\|I_s f \mathbb{1}_{\mathbb{B}(0,2)}\|_2$. Let $q = \frac{2}{1-\gamma/d} > 2$ and let $\|\cdot\|_{q,w}$ denote the weak q-norm. Define the distribution function $\lambda(t) = m\{x \in \mathbb{R}^d : |I_s f(x)\mathbb{1}_{\mathbb{B}(0,2)}| > t\}$ where m is the Lebesgue measure on $\mathbb{R}^d$. Because $\lambda(t) \le \min\left(t^{-q}\|I_s f \mathbb{1}_{\mathbb{B}(0,2)}\|_{q,w}^q, m(\mathbb{B}(0,2))\right)$ for any $t \ge 0$, we have

$$\|I_s f \mathbb{1}_{\mathbb{B}(0,2)}\|_2 = \left(2\int_0^\infty t\lambda(t)\mathrm{d}t\right)^{\frac{1}{2}} \le C_2\|I_s f \mathbb{1}_{\mathbb{B}(0,2)}\|_{q,w}$$

where $C_2(d,\gamma) = (\frac{q}{q-2})^{\frac{1}{2}} m(\mathbb{B}(0,2))^{\frac{1}{2}-\frac{1}{q}}$. By the Hardy-Littlewood-Sobolev lemma [185, Theorem V.1],

$$\|I_s f\|_{q,w} \lesssim \|f\|_1 = 2\epsilon,$$

where we again discard a constant depending on d and γ. Combining these inequalities together, we get

$$\|I_s f\|_2 \leq \|I_s f \mathbb{1}_{\mathbb{B}(0,2)^c}\|_2 + \|I_s f \mathbb{1}_{\mathbb{B}(0,2)}\|_2 \leq C(d,\gamma)\epsilon$$

for some constant $C(d,\gamma)$.

Finally, if either μ or ν does not have a density, we pick a positive mollifier $\phi \in C_c^\infty(\mathbb{R}^d)$ such that $\operatorname{supp}(\phi) \subseteq \mathbb{B}(0,1)$, $\phi \geq 0$, and $\int \phi = 1$ and set $\phi_\eta(x) = \eta^{-d}\phi(x-\eta)$ for $\eta > 0$. Consider $\mu_\eta = \mu * \phi_\eta$ and $\nu_\eta = \nu * \phi_\eta$, which are both absolutely continuous and supported on $\mathbb{B}(0,1+\eta)$. By the first part of our proof so far we get $\mathcal{E}_\gamma(\mu_\eta,\nu_\eta) \leq C(d,\gamma)(1+\eta)^{\gamma/2}\mathsf{TV}(\mu_\eta,\nu_\eta)$ by a simple argument that rescales the two mollified distributions to be supported on the unit ball. It is well known that $\mu_\eta \to \mu$ and $\nu_\eta \to \nu$ as $\eta \to 0$ in a weak sense, thus by the definition of $\mathcal{E}_\gamma(\mu_\eta,\nu_\eta)$ given in (5.2.1), we obtain $\mathcal{E}_\gamma(\mu_\eta,\nu_\eta) \to \mathcal{E}_\gamma(\mu,\nu)$ as $\eta \to 0$. Moreover, we have $\mathsf{TV}(\mu_\eta,\nu_\eta) \leq \mathsf{TV}(\mu,\nu)$ by the data processing inequality, which concludes the proof. □

5.3.2 Lower Bound — Smooth Distributions And Gaussian Mixtures

In Section 5.3.1 we showed that the energy distance is upper bounded by total variation for compactly supported measures. In this section we look at the reverse direction, namely, we aim to lower bound the energy distance by total variation.

Theorem 5.3.2. *For any $\beta > 0$, $d \geq 1$ and $C > 0$, there exists a finite constant C_1 so that*

$$\sqrt{\gamma(2-\gamma)}\mathsf{TV}(\mu,\nu)^{\frac{2\beta+d+\gamma}{2\beta}} \leq C_1\mathcal{E}_\gamma(\mu,\nu) \tag{5.3.1}$$

for any $\mu,\nu \in \mathcal{P}_S(\beta,d,C)$ and $\gamma \in (0,2)$. Similarly, for any $d \geq 1$ there exists a finite constant C_2 such that

$$\frac{\sqrt{\gamma(2-\gamma)}\mathsf{TV}(\mu,\nu)}{\log(3+1/\mathsf{TV}(\mu,\nu))^{\frac{2d+\gamma}{4}}} \leq C_2\mathcal{E}_\gamma(\mu,\nu) \tag{5.3.2}$$

for every $\mu,\nu \in \mathcal{P}_G(d)$ and $\gamma \in (0,2)$.

Proof. Abusing notation, identify μ and ν with their Lebesgue densities. The argument proceeds trough a chain of inequalities:

1. Bound TV by the L^2 distance between densities.
2. Apply Parseval's Theorem to pass to Fourier space.
3. Apply Hölder's inequality with well-chosen exponents.

Proof of (5.3.1). Jensen's inequality implies that

$$2\mathsf{TV}(\mu,\nu) = \|\mu-\nu\|_1 \le \sqrt{\mathrm{vol}(\mathbb{B}(0,1))}\|\mu-\nu\|_2 \lesssim \|\mu-\nu\|_2,$$

where vol denotes volume and we discard dimension-dependent constants. This completes the first step of our proof. For the second step note that $\mu,\nu \in L^2(\mathbb{R}^d)$ and we may apply Parseval's theorem to obtain

$$\|\mu-\nu\|_2^2 = \frac{1}{(2\pi)^d}\|\widehat{\mu}-\widehat{\nu}\|_2^2.$$

For arbitrary $\varphi > 0$ and $r \in [1,\infty]$, Hölder's inequality with exponents $\frac{1}{r}+\frac{1}{r^*}=1$ implies that

$$\begin{aligned}\|\widehat{\mu}-\widehat{\nu}\|_2^2 &= \int_{\mathbb{R}^d} |\widehat{\mu}(\omega)-\widehat{\mu}(\omega)|^2 \frac{\|\omega\|^\varphi}{\|\omega\|^\varphi}\mathrm{d}\omega \\ &\le \left(\int_{\mathbb{R}^d} |\widehat{\mu}(\omega)-\widehat{\nu}(\omega)|^2\|\omega\|^{\varphi r}\mathrm{d}\omega\right)^{1/r}\left(\int_{\mathbb{R}^d}\frac{|\widehat{\mu}(\omega)-\widehat{\nu}(\omega)|^2}{\|\omega\|^{\varphi r^*}}\mathrm{d}\omega\right)^{1/r^*}.\end{aligned} \tag{5.3.3}$$

Now, we choose φ and r to satisfy

$$\begin{aligned}\varphi r &= 2\beta \\ \varphi r^* &= d+\gamma.\end{aligned}$$

The first equation ensures that the first integral term is bounded by $\|\mu-\nu\|_{\beta,2}^{2/r}$, which is assumed to be at most a d,β dependent constant. The second equation ensures that the second integral term is equal to $(\mathcal{E}_\gamma(\mu,\nu)^2/F_\gamma(d))^{1/r^*}$ by Proposition 5.2.3. The solution to this system of equations is given by $r^* = (2\beta+d+\gamma)/(2\beta)$ and $\varphi = 2\beta\cdot\frac{d+\gamma}{2\beta+d+\gamma}$. Note that clearly $\varphi > 0$ and $r^* \ge 1$. Thus, after rearrangement and using that $F_d(\gamma) = \Theta(\gamma(2-\gamma))$ up to a dimension dependent constant, we obtain

$$\sqrt{\gamma(2-\gamma)}\|\widehat{\mu}-\widehat{\nu}\|_2^{\frac{2\beta+d+\gamma}{2\beta}} \le C_1\mathcal{E}_\gamma(\mu,\nu),$$

for a finite constant $C_1 = C_1(d,\beta)$, concluding the proof.

Proof of (5.3.2). We write $C(d) \in (0,\infty)$ for a dimension dependent constant that may change from line to line. The outline of the argument is analogous to the above, with the additional step of having to bound the $(\beta,2)$-Sobolev norm of the Gaussian density as $\beta\to\infty$ for which we rely on Lemma D.1.6. Let μ and ν have densities $p*\phi$ and $q*\phi$, where ϕ is the density of $\mathcal{N}(0,I_d)$. Writing $f = (p-q)*\phi$, we can extend the proof of [116, Theorem 22] to multiple dimensions to find, for any $R > 2$, that

$$\begin{aligned}2\mathsf{TV}(\mu,\nu) = \|\mu-\nu\|_1 &= \int_{\|x\|\le R}|(f*\phi)(x)|\,\mathrm{d}x + \int_{\|x\|>R}\left|\int_{\mathbb{R}^d}\phi(x-y)\mathrm{d}f(y)\right|\mathrm{d}x \\ &\le \sqrt{\mathrm{vol}_d(\mathbb{B}(0,R))}\sqrt{\int_{\|x\|\le R}|(f*\phi)(x)|^2\mathrm{d}x} + \int_{\|x\|>R}\exp(-\|x\|^2/8)\mathrm{d}x \\ &\le C(d)\Big(R^{d/2}\|\mu-\nu\|_2 + \exp(-\Omega(R^2))\Big),\end{aligned}$$

where the second line uses that $\operatorname{supp}(f) \subseteq \mathbb{B}(0,1)$. Taking $R \asymp \sqrt{\log(3+1/\|\mu-\nu\|_2)}$ we obtain the inequality

$$\mathsf{TV}(\mu,\nu) \le C(d)\|\mu-\nu\|_2 \log(3+1/\|\mu-\nu\|_2)^{d/4}. \tag{5.3.4}$$

By Hölder's inequality we obtain

$$\begin{aligned}\|\widehat{f}\|_2 &\le \||\omega|^\beta \widehat{f}(\omega)\|_2^{\frac{d+\gamma}{2\beta+d+\gamma}} \left\| \frac{\widehat{f}(\omega)}{\|\omega\|^{\frac{d+\gamma}{2}}} \right\|_2^{\frac{2\beta}{2\beta+d+\gamma}} \\ &= \||\omega|^\beta \widehat{f}(\omega)\|_2^{\frac{d+\gamma}{2\beta+d+\gamma}} \cdot \mathcal{E}_\gamma(\mu,\nu)^{\frac{2\beta}{2\beta+d+\gamma}} \cdot F_\gamma(d)^{-\frac{\beta}{2\beta+d+\gamma}}\end{aligned}$$

by Proposition 5.2.3. Using that $|\widehat{f}| \le |\widehat{\phi}|$ and applying Lemma D.1.6, for $\beta \ge 1$ we get

$$F_\gamma(d)^{\frac{\beta}{2\beta+d+\gamma}} \|\widehat{f}\|_2 \le \mathcal{E}_\gamma(\mu,\nu)^{\frac{2\beta}{2\beta+d+\gamma}} \left(\frac{5\pi^{d/2}}{\Gamma(d/2)} \left(\frac{2\beta+d}{2e} \right)^{\frac{2\beta+d-1}{2}} \right)^{\frac{d+\gamma}{2(2\beta+d+\gamma)}}.$$

Rearranging and using Parseval's Theorem, we get

$$\mathcal{E}_\gamma(\mu,\nu) \ge C(d)\sqrt{\gamma(2-\gamma)}\|f\|_2 \frac{\|f\|_2^{\frac{d+\gamma}{2\beta}}}{\left(\frac{2\beta+d}{2e}\right)^{\frac{(d+\gamma)(2\beta+d-1)}{8\beta}}}$$

for some d-dependent, albeit exponential, constant $C(d) > 0$. Plugging in $\beta = \log(3+1/\|f\|_2)$ and assuming that $\|f\|_2$ is small enough in terms of d, we obtain

$$\mathcal{E}_\gamma(\mu,\nu) \ge \frac{C(d)\sqrt{\gamma(2-\gamma)}\|f\|_2}{\log(3+1/\|f\|_2)^{\frac{d+\gamma}{4}}} \ge \frac{C(d)\sqrt{\gamma(2-\gamma)}\mathsf{TV}(\mu,\nu)}{\log(3+1/\mathsf{TV}(\mu,\nu))^{\frac{2d+\gamma}{4}}}, \tag{5.3.5}$$

where the second inequality uses (5.3.4) and Lemma D.1.2. □

Theorem 5.3.2 is our main technical result, which shows that $\mathcal{E}_\gamma$ is lower bounded by a polynomial of the total variation distance for both the smooth distribution class $\mathcal{P}_S$ and Gaussian mixtures $\mathcal{P}_G$. Note also that in one dimension, (5.3.1) follows from the Gagliardo–Nirenberg-Sobolev interpolation inequality. However, to our knowledge, the inequality is new for $d > 1$. As for the tightness of Theorem 5.3.2, we manage to prove that this inequality is the best possible for $\mathcal{P}_S$ in one dimension, and best possible up to a poly-logarithmic factor in dimension 2 and above.

Proposition 5.3.3. *For any $\beta > 0$, $d \ge 1$, $\gamma \in (0,2)$ and $C > 0$ satisfying Assumption 3, there exists a finite constant C_1 so that for any value of $\epsilon \in (0,1)$, there exist $\mu_\epsilon, \nu_\epsilon \in \mathcal{P}_S(\beta,d,C)$ such that $\mathsf{TV}(\mu_\epsilon,\nu_\epsilon)/\epsilon \in (1/C_1, C_1)$ and*

$$\mathcal{E}_\gamma(\mu_\epsilon,\nu_\epsilon) \le C_1 \mathsf{TV}(\mu_\epsilon,\nu_\epsilon)^{\frac{2\beta+d+\gamma}{2\beta}} \log\left(3+\frac{1}{\mathsf{TV}(\mu_\epsilon,\nu_\epsilon)}\right)^{d-1}.$$

In the special case $\gamma = 1$ we obtain an even stronger notion of tightness.

Proposition 5.3.4. *When $\gamma = 1$ we may replace $\mathcal{E}_1$ by $\overline{d_H}$ in Proposition 5.3.3.*

Proposition 5.3.4 is an improvement over Proposition 5.3.3 due to the inequality $d_H \lesssim \overline{d_H}$ over the class $\mathcal{P}_S(\beta, d, C)$, which follows from Proposition 5.2.1. It shows also that our construction has the property that there does not exist any halfspace that separates μ and ν better than our bounds suggest.

The proofs of both results are presented in Appendix D.4. The general idea is to saturate Hölder's inequality in (5.3.3), for which the Fourier transform of $f = \frac{\mathrm{d}\nu}{\mathrm{d}x} - \frac{\mathrm{d}\mu}{\mathrm{d}x}$ should be supported on a sphere. However, such f clearly cannot be compactly supported. Thus the actual construction is to multiply the Fourier inverse of the uniform measure on a sphere with a compactly supported mollifier. In $d > 1$ the mollifier that we require must have super-polynomial Fourier spectrum decay, for which we use the recent construction in [47].

5.3.3 Lower Bound — Discrete Distributions

Suppose we have two discrete distributions that are supported on a common, finite set of size k. One way to measure the energy distance between them would be to identify their support with the set $\{1, 2, \ldots, k\}$, thereby embedding the two distributions in $\mathbb{R}$, and applying the one-dimensional energy distance.

While the above approach seems reasonable, it is entirely arbitrary. Indeed, there might not be a natural ordering of the support; moreover, why should one choose the integers between 1 and k instead of, say, the set $\{1, 2, 4, \ldots, 2^k\}$? The total variation distance does not suffer from such ambiguities, and it is unclear how our choice of embedding affects the relationship to TV. The following result attacks precisely this question.

Theorem 5.3.5. *Let μ and ν be probability distributions supported on the set $\{x_1, \ldots, x_k\} \subseteq \mathbb{R}^d$ and let $\delta = \min_{i \neq j} \|x_i - x_j\|$. Then there exists a universal constant $C > 0$ such that*

$$\mathcal{E}_1^2(\mu, \nu) \geq \frac{C\delta}{k\sqrt{d}} \mathsf{TV}^2(\mu, \nu).$$

Proof. Let $\mu = \sum_{i=1}^k \mu_i \delta_{x_i}$ and $\nu = \sum_{i=1}^k \nu_i \delta_{x_i}$. Then, by [15, Theorem 1] we have

$$\mathcal{E}_1^2(\mu, \nu) = -\sum_{i,j} (\mu_i - \nu_i)(\mu_j - \nu_j)\|x_i - x_j\| \geq \frac{C\delta}{\sqrt{d}} \sum_{i=1}^k (\mu_i - \nu_i)^2 \geq \frac{C\delta \mathsf{TV}^2(\mu, \nu)}{k\sqrt{d}}$$

as required.

□

Remark 19. *Similar results can be proved for the generalized energy distance $\mathcal{E}_\gamma$, using e.g. the work [153]. However, to the best of our knowledge, these estimates degrade significantly in the dimension d in contrast with [15].*

Notice that by our discussion above, the support set $\{x_1, \ldots, x_k\}$ in Theorem 5.3.5 is arbitrary and may be chosen by us. Since the scale of the supporting points $x_1, \ldots, x_k$ is statistically irrelevant, we remove this ambiguity by restricting the points to lie in the unit

ball, i.e. requiring that $\max_i \|x_i\| \leq 1$. We see now that the comparison between $\mathcal{E}_1$ and TV *improves* as $\delta/\sqrt{d}$ grows. Given a fixed value of δ, we want to make the dimension d of our embedding as low as possible, which means that the points $x_1, \ldots, x_k$ should form a large δ-packing of the d-dimensional unit ball. Due to well known bounds on the packing number of the Euclidean ball, it follows that the best one can hope for is

$$\log(k) \asymp d \log(1/\delta).$$

Maximizing $\delta/\sqrt{d}$ subject to this constraint yields the choice $d = \Theta(\log(k))$ and $\delta = \Theta(1)$. This gives us the following corollary.

Corollary 5.3.6. *There exists a universal constant $C \in (0, \infty)$ such that for any $k \geq 1$ there exists a set of points $x_1, \ldots, x_k \in \mathbb{R}^{\lceil C \log(k) \rceil}$ with $\max_i \|x_i\| \leq 1$ such that*

$$\mathcal{E}_1\left(\sum_{i=1}^k \mu_i \delta_{x_i}, \sum_{i=1}^k \nu_i \delta_{x_i}\right) \geq \frac{\mathsf{TV}(\mu, \nu)}{C\sqrt{k}\sqrt[4]{\log(k)}}$$

for any two probability mass functions $\mu = (\mu_1, \ldots, \mu_k)$ and $\nu = (\nu_1, \ldots, \nu_k)$.

The question arises how the set of points $x_1, \ldots, x_k$ in Corollary 5.3.6 should be constructed. One solution is to use an error correcting code (ECC), whereby we take the x_i to be the codewords of an ECC on the scaled hypercube $\frac{1}{\sqrt{d}}\{\pm 1\}^d$ for some dimension d (known as "blocklength" in this context). An ECC is *asymptotically good* if the message length $\log(k)$ is linear in the blocklength d, that is $d \asymp \log(k)$, and if the minimum Hamming distance between any two codewords is $\Theta(d)$, which translates precisely into $\delta = \min_{i \neq j} \|x_i - x_j\| \asymp 1$. Many explicit constructions of asymptotically good error correcting codes exist, see [120] for one such example, and random codes are almost surely good [17]. Clearly the better the code is, the better the constants we obtain in Corollary 5.3.6.

Remark 20. *One interesting consequence of Corollary 5.3.6 and the preceeding discussion is the following: given a categorical feature with k possible values, the perceptron may obtain better performance by identifying each category with the codewords $x_1, \ldots, x_k$ of an ECC instead of the standard one-hot encoding.*

5.4 Density Estimation

In this section we apply what we've learnt about the generalized energy distance and the perceptron discrepancy in prior sections, and analyze multiple problems related to density estimation.

5.4.1 Estimating Smooth Distributions and Gaussian Mixtures

Suppose that $X_1, \ldots, X_n \overset{iid}{\sim} \nu$ for some probability distribution ν on $\mathbb{R}^d$. Given a class of "generator" distributions $\mathcal{G}$ and $\gamma \in (0, 2)$, define the minimum-$\mathcal{E}_\gamma$ estimator as

$$\tilde{\nu}_\gamma \in \arg\min_{\nu' \in \mathcal{G}} \mathcal{E}_\gamma(\nu', \nu_n), \tag{5.4.1}$$

where $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. Note that $\tilde{\nu}_\gamma$ does not quite agree with our definition of $\tilde{\nu}_\gamma$ in (5.1.2), because the $\gamma = 1$ case minimizes the *average* halfspace distance $d_H \asymp \mathcal{E}_1$ and not the perceptron discrepancy $\overline{d_H}$. The following two results bound the performance of $\tilde{\nu}$ as defined in (5.4.1), as an estimator of ν for the smooth density class $\mathcal{P}_S$ as well as the Gaussian mixture class $\mathcal{P}_G$. In Section 5.4.2 we present the adapation of these to $\overline{d_H}$, thereby proving Theorem 5.1.1.

Theorem 5.4.1. *Let $\tilde{\nu}_\gamma$ be the estimator defined in (5.4.1). For any $\beta > 0$, $d \geq 1$ and $C > 0$, there exists a finite constant C_1 so that*

$$\sup_{\nu \in \mathcal{P}_S(\beta,d,C)} \mathbb{E}\mathsf{TV}(\tilde{\nu}_\gamma, \nu) \leq C_1 (n\gamma(2-\gamma))^{-\frac{\beta}{2\beta+d+\gamma}} \tag{5.4.2}$$

holds for $\mathcal{G} = \mathcal{P}_S(\beta, d, C)$ and any $\gamma \in (0,2)$. Similarly, for any $d \geq 1$ there is a finite constant C_2 such that

$$\sup_{\nu \in \mathcal{P}_G(d)} \mathbb{E}\mathsf{TV}(\tilde{\nu}_\gamma, \nu) \leq C_2 \phi(n\gamma(2-\gamma)) \tag{5.4.3}$$

holds for $\mathcal{G} = \mathcal{P}_G(d)$ and any $\gamma \in (0,2)$, where $\phi(x) = \frac{\log(3+x)^{\frac{2d+\gamma}{4}}}{\sqrt{x}}$.

Proof. Let us focus on the case $\mathcal{G} = \mathcal{P}_S(\beta, d, C)$ first and let $t = \frac{2\beta+d+\gamma}{2\beta}$. The inequality $\mathcal{E}_\gamma(\tilde{\nu}_\gamma, \nu_n) \leq \mathcal{E}_\gamma(\nu, \nu_n)$ holds almost surely by the definition of $\tilde{\nu}_\gamma$. Writing $C_1 = C_1(\beta, d, C)$ for a finite constant that we relabel freely, the first claim is substantiated by the chain of inequalities

$$\begin{aligned}
\mathbb{E}\mathsf{TV}(\tilde{\nu}_\gamma, \nu) &\overset{\text{Thm. 5.3.2}}{\leq} \mathbb{E}\Big[\left(C_1 \frac{\mathcal{E}_\gamma(\tilde{\nu}_\gamma, \nu)}{\sqrt{\gamma(2-\gamma)}}\right)^{1/t}\Big] \\
&\overset{\Delta-\text{ineq.}}{\leq} \mathbb{E}\Big[\left(C_1 \frac{\mathcal{E}_\gamma(\nu, \nu_n) + \mathcal{E}_\gamma(\tilde{\nu}_\gamma, \nu_n)}{\sqrt{\gamma(2-\gamma)}}\right)^{1/t}\Big] \\
&\overset{\text{Eq. (5.4.1)}}{\leq} \mathbb{E}\Big[\left(2C_1 \frac{\mathcal{E}_\gamma(\nu, \nu_n)}{\sqrt{\gamma(2-\gamma)}}\right)^{1/t}\Big] \\
&\overset{\text{Jensen's}}{\leq} \left(2C_1 \frac{\mathbb{E}\mathcal{E}_\gamma(\nu, \nu_n)}{\sqrt{\gamma(2-\gamma)}}\right)^{1/t} \\
&\overset{\text{Lem. 5.2.4}}{\leq} \left(\frac{n\gamma(2-\gamma)}{8C_1^2}\right)^{-1/2t}.
\end{aligned}$$

The result for $\mathcal{G} = \mathcal{P}_G$ follows analogously. Define the function $r(x) = x\sqrt{\gamma(2-\gamma)}/\log(3 + 1/x)^{\frac{2d+\gamma}{4}}$. One can check by direct calculation that r is strictly increasing and convex on $\mathbb{R}_+$. As a consequence, its inverse r^{-1} is strictly increasing and concave. Let C_2 be a d-dependent finite constant which we relabel repeatedly. Similarly to the case of smooth distributions

covered above, using Theorem 5.3.2 and Jensen's inequality we obtain the chain of inequalities

$$\begin{aligned}\mathbb{E}\mathsf{TV}(\tilde{\nu}_\gamma, \nu) &\leq \mathbb{E}\left[r^{-1}\left(C_2\frac{\mathcal{E}_\gamma(\tilde{\nu}_\gamma, \nu)}{\sqrt{\gamma(2-\gamma)}}\right)\right] \\ &\leq r^{-1}\left(C_2\frac{\mathbb{E}\mathcal{E}_\gamma(\tilde{\nu}_\gamma, \nu)}{\sqrt{\gamma(2-\gamma)}}\right) \\ &\leq r^{-1}\left(2C_2\frac{\mathbb{E}\mathcal{E}_\gamma(\nu, \nu_n)}{\sqrt{\gamma(2-\gamma)}}\right) \\ &\leq r^{-1}\left(C_2(n\gamma(2-\gamma))^{-1/2}\right).\end{aligned}$$

The conclusion follows by Lemma D.1.2. □

Notice that the rate of estimation of the minimum $\mathcal{E}_\gamma$ density estimator improves as $\gamma \downarrow 0$, and in fact seems to approach the optimum. However, simultaneously, the "effective sample size" $n\gamma$ shrinks. The best trade-off that we can derive is the following.

Corollary 5.4.2. *The rate in (5.4.2) (resp. (5.4.3)) can be improved to* $(\log(n)/n)^{\beta/(2\beta+d)}$ *(resp.* $\log(n)^{d/4}\sqrt{\log\log n}/\sqrt{n}$*) by adaptively setting* $\gamma = \log(n)^{-1}$ *(resp.* $\gamma = \log\log(n)^{-1}$*).*

5.4.2 Proof of Theorem 5.1.1 and Theorem 5.1.2

We already have everything needed to deduce Theorem 5.1.2. Since it is an exercise in combining results, we simply list the required steps:

1. Use Theorem 5.3.2 to get a comparison between TV and $\mathcal{E}_\gamma$.
2. Set $\gamma = 1$ and use Proposition 5.2.2 to get the equivalence between $\mathcal{E}_1$ and d_H.
3. Use Proposition 5.2.1 to get a comparison between d_H and $\overline{d_H}$.

Turning to the proof of Theorem 5.1.1, we find that it is completely analogous to the proof of Theorem 5.4.1, with the only difference being that we can no longer rely on Lemma 5.2.4 to show that the distance between empirical and population measures decays at the parametric rate, as the latter applies to $\mathcal{E}_\gamma$ instead of $\overline{d_H}$. However, the corresponding result for $\overline{d_H}$ is well known.

Lemma 5.4.3. *Let* ν *be a probability distribution on* $\mathbb{R}^d$ *and* $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ *for i.i.d. observations* $X_i \sim \nu$*. Then, for a finite universal constant* C*,*

$$\mathbb{E}\overline{d_H}(\nu, \nu_n) \leq C\sqrt{\frac{d}{n}}.$$

Proof. Follows for example from [201, p. 8.3.23] and the fact that $\mathcal{D}_a$, the family of halfspace indicators, has VC dimension $d+1$. □

With Lemma 5.4.3 in hand, completing the argument is straightforward: To deduce Theorem 5.1.1 follow the same steps as in the proof of Theorem 5.4.1, except use Theorem 5.1.2 and Lemma 5.4.3 in place of Theorem 5.3.2 and Lemma 5.2.4 respectively.

5.4.3 Estimating Discrete Distributions

In many practical machine learning tasks the data is discrete, albeit on a large alphabet $[k] = \{1, 2, \dots, k\}$: for example, in recommender systems the alphabet could be all possible ads, products or articles. A common idea to apply modern learning pipelines to such data is to use an embedding $E : [k] \to \mathbb{R}^d$, with "one-hot" encoding ($d = k$) being the most popular choice. After such an embedding, the data is effectively made "continuous" and the density estimation methods as discussed previously can be applied. Can such an approach be good in the sense of minimax estimation guarantees? We answer this question positively in this section, provided that embedding E comes from an error-correcting code.

Let $\mathcal{P}_k$ denote the set of all probability distributions on the set $[k]$. Suppose we observe an i.i.d. sample $X_1, \dots, X_n \sim \nu$ from some unknown distribution $\nu \in \mathcal{P}_k$. The problem of estimating ν is effectively trivial: the empirical distribution provides a minimax optimal estimator. Indeed, it is a folklore fact, see for example [35, Theorem 1] or [168, Exc. VI.8], that the optimal rate of estimation is given by

$$\sup_{\nu \in \mathcal{P}_k} \mathbb{E}\mathsf{TV}^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{X_i}, \nu\right) \asymp \min\left\{\frac{k}{n}, 1\right\}. \tag{5.4.4}$$

Recall from Section 5.3.3 that we may choose to embed the alphabet $[k]$ into some higher dimensional Euclidean space. Given distinct points $x_1, \dots, x_k \in \mathbb{R}^d$ for some $d \geq 1$, we can identify any distribution $\mu \in \mathcal{P}_k$ with the probability distribution $\sum_{i=1}^k \mu_i \delta_{x_i}$, where μ_i is the mass that μ puts on $i \in [k]$.

Theorem 5.4.4. *There exists a universal constant $C < \infty$ with the following property. For any alphabet size k there exist embedding points $a_1, \dots, a_k \in \mathbb{R}^{\lceil C \log(k) \rceil}$ such that given an i.i.d. sample $X_1, \dots, X_n \sim \nu$ from an unknown $\nu \in \mathcal{P}_k$, any estimator $\tilde{\nu} \in \mathcal{P}_k$ that satisfies*

$$\mathcal{E}_1^2\left(\sum_{i=1}^k \tilde{\nu}_i \delta_{a_i}, \frac{1}{n}\sum_{i=1}^n \delta_{a_{X_i}}\right) \leq \frac{c}{n} \tag{5.4.5}$$

enjoys the performance guarantee

$$\sup_{\nu \in \mathcal{P}_k} \mathbb{E}\mathsf{TV}^2(\tilde{\nu}, \nu) \leq C \min\left\{(c+1)\frac{k\sqrt{\log(k)}}{n}, 1\right\}. \tag{5.4.6}$$

Moreover, we may replace $\mathcal{E}_1$ by $\overline{d_H}$ in (5.4.5) and the result (5.4.6) remains true with $\sqrt{\log(k)}$ replaced by $\log(k)$.

Proof. Let $a_1, \dots, a_k \in \mathbb{R}^d$ be the points defined in Corollary 5.3.6 (relabeled from $x_1, \dots, x_k$ for clarity) so that $d \asymp \log(k)$. By the triangle inequality we have

$$\begin{aligned}\mathbb{E}\mathcal{E}_1^2\left(\sum_{i=1}^k \tilde{\nu}_i \delta_{a_i}, \sum_{i=1}^k \nu_i \delta_{a_i}\right) &\leq 2\mathbb{E}\mathcal{E}_1^2\left(\sum_{i=1}^k \tilde{\nu}_i \delta_{a_i}, \frac{1}{n}\sum_{i=1}^n \delta_{a_{X_i}}\right) + 2\mathbb{E}\mathcal{E}_1^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{a_{X_i}}, \sum_{i=1}^k \nu_i \delta_{a_i}\right) \\ &\overset{\text{Lemma 5.2.4}}{\lesssim} \frac{c + \max_i \|a_i\|}{n} \lesssim \frac{c+1}{n}.\end{aligned}$$

By Corollary 5.3.6, the definition of $\tilde{\nu}$ and the triangle inequality it follows that

$$\mathbb{E}\mathsf{TV}^2(\tilde{\nu},\nu) \lesssim \frac{k\sqrt{d}(c+1)}{n} \asymp \frac{k\sqrt{\log(k)}(c+1)}{n}.$$

Noting the trivial fact that $\mathsf{TV} \leq 1$ completes the proof of the first claim.

Suppose now that we replace $\mathcal{E}_1$ by $\overline{d_H}$ in the definition of $\tilde{\nu}$. The proof follows analogously, using the chain of inequalities

$$\frac{\mathsf{TV}}{\sqrt{k}\sqrt[4]{d}} \overset{\text{Cor. 5.3.6}}{\lesssim} \mathcal{E}_1 \overset{\text{Prop. 5.2.2}}{\asymp} \frac{\sqrt{\Gamma\left(\frac{d+1}{2}\right)}}{\pi^{(d-1)/4}} d_H \overset{\max_i \|a_i\| \leq 1}{\lesssim} \overline{d_H},$$

and Lemma 5.4.3 in place of Lemma 5.2.4, which is where we loose the $\sqrt{d} \asymp \sqrt{\log(k)}$ factor. □

As we explained, the empirical distribution $\tilde{\nu} = \frac{1}{n}\sum_{i=1} \delta_{X_i}$ achieves optimality in (5.4.4), and clearly also achieves $c = 0$ in (5.4.5) i.e. minimizes the empirical risk globally. The point of Theorem 5.4.4 is to show that approximate minimizers, such as those found via SGD, are also nearly minimax optimal.

Estimating Hölder Smooth Densities

Theorem 5.4.4 has interesting implications for density estimation over the class of distributions on the cube $[0,1]^d$ with uniformly bounded derivatives up to order $\underline{\beta} =: \lceil \beta - 1 \rceil$ and $(\beta - \underline{\beta})$-Hölder continuous $\underline{\beta}^{th}$ derivative; call such distributions simply β-Hölder smooth.[5] Writing B_j for the cube with center $(j - \frac{1}{2})\epsilon^{1/\beta}$ and sidelength $\epsilon^{1/\beta}$ where $j \in \{1, \ldots, \epsilon^{-1/\beta}\}^d$, it is known that

$$\sum_j \left| \int_{B_j} (f(x) - g(x)) \mathrm{d}x \right| = c \int_{[0,1]^d} |f(x) - g(x)| \mathrm{d}x + O(\epsilon)$$

for any β-Hölder smooth densities f, g and an ϵ-independent constant $c > 0$, see for example [10, Lemma 7.2] or [113, Proposition 2.16]. In other words, discretizing such distributions using a regular grid with $\Omega(\epsilon^{-d/\beta})$ cells maintains total-variation distances up to an additive $O(\epsilon)$ error.

Now, consider a 'multilayer perceptron', that is, a fully connected multilayer neural network with activations given by $x \mapsto \mathbb{1}\{x \geq 0\}$. Such a multilayer network with large enough hidden layers can in principle implement the discretization described above, and embed the $\epsilon^{-d/\beta}$ cells as an error correcting code. Thus, due to Theorem 5.4.4, the ERM density estimator (5.1.2) would achieve the minimax optimal density estimation rate $n^{-\beta/(2\beta+d)}$ over β-Hölder smooth densities up to polylog factors provided the discriminator class $\mathcal{D}$ includes the aforementioned multilayer perceptron and has VC-dimension at most polylog in $1/\epsilon$. This observation essentially generalizes Theorem 5.1.1, which shows that if the discriminator class includes only the *single* layer perceptron then the best possible minimax rate is $n^{-\beta/(2\beta+d+1)}$.

[5] Note that this class is not the same as $\mathcal{P}_S$, although related.

5.4.4 A Stopping Criterion for Smooth Density Estimation

As a corollary to our results, we propose a stopping criterion for training density estimators. Before doing so, let us record a result about the concentration properties of the empirical energy distance about its expectation.

Lemma 5.4.5. *Let ν be supported on a compact subset $\Omega \subseteq \mathbb{R}^d$, and let ν_n be its empirical measure based on n i.i.d. observations. For every $\gamma \in (0,2)$ there exists a constant $C_1 \in (0,\infty)$ such that*

$$\mathbb{P}\left(\mathcal{E}_\gamma(\nu,\nu_n) \geq \frac{C_1}{\sqrt{n}} + t\right) \leq 2\exp\left(-\frac{nt^2}{C_1}\right).$$

In other words, $\mathcal{E}_\gamma(\nu,\nu_n)$ is $O(1/n)$-sub-Gaussian.

Proof. Recall the MMD formulation of the generalized energy distance from Section 5.2.3. The corresponding kernel is given by $k_\gamma(x,y) = \|x\|^\gamma + \|y\|^\gamma - \|x-y\|^\gamma$. Clearly

$$\sup_{x,x',y,y' \in \mathrm{supp}(\nu)} (k_\gamma(x,y) - k_\gamma(x',y')) \lesssim \mathrm{diam}(\Omega)^\gamma.$$

Therefore, by McDiarmid's inequality we know that $\mathcal{E}_\gamma(\nu,\nu_n)$ is $O(1/n)$-subGaussian (note we don't track constants depending on Ω here). From Lemma 5.2.4 we know that $\mathbb{E}\mathcal{E}_\gamma(\nu,\nu_n) \lesssim 1/\sqrt{n}$, and the conclusion follows. □

Consider the following scenario: we have i.i.d. training data $X_1,\dots,X_n$ from some distribution ν and we are training an arbitrary generative model to estimate ν. Suppose that this training process gives us a sequence of density estimators $\{\mu_k\}_{k\geq 1}$, which could be the result of, say, subsequent gradient descent steps on our parametric class of generators. Is there any way to figure out after how many steps K we may stop the training process? In other words, can we identify a value of K such that $\mathsf{TV}(\nu,\mu_K)$ is guaranteed to be less than some threshold with probability $1-\delta$? Note that an additional difficulty here is that our generative model for μ_k is able to generate the samples from μ_k but otherwise gives us no other access to μ_k. The fast (dimension-free) concentration properties of $\mathcal{E}_\gamma$ and the minimax optimality guarantees of its minimizer (whenever ν is smooth) make it an excellent choice for such a stopping criteria.

Let $\nu \in \mathcal{P}_S(\beta,d,C)$ and let ν_n be its empirical version based on the n i.i.d. observations. Assume further that $\{\mu_k\}_{k\geq 1} \subseteq \mathcal{P}_S(\beta,d,C)$ is a sequence of density estimators based on the sample $X_1,\dots,X_n$. Finally, given the training sample $(X_1,\dots,X_n)$, for each k let $\mu_{k,m_k} = \frac{1}{m_k}\sum_{i=1}^{m_k} \delta_{X_i^{(k)}}$ be the empirical distribution of the sample $(X_1^{(k)},\dots,X_{m_k}^{(k)}) \sim \mu_k^{\otimes m_k}$.

Proposition 5.4.6. *For any $\beta > 0, d \geq 1$ and $\gamma \in (0,2)$ there exists a constant $c \in (0,\infty)$ such that*

$$\mathbb{P}\left(\mathsf{TV}(\mu_k,\nu) \leq c\left(\sqrt{\frac{\log(1/\delta)}{n}} + \mathcal{E}_\gamma(\mu_{k,m_k},\nu_n)\right)^{\frac{2\beta}{2\beta+d+\gamma}},\ \forall\, k \geq 1\right) \geq 1-2\delta$$

provided we take $m_k = cn\log(k^2/\delta)/\log(1/\delta)$.

Proof. Let $c = C_1$ where C_1 is as in Lemma 5.4.5 and fix $\delta \in (0,1)$. Define the event $A = \left\{\mathcal{E}_\gamma(\nu,\nu_n) \geq \frac{c}{\sqrt{n}} + \sqrt{\frac{c\log(2/\delta)}{n}}\right\}$ and similarly

$$A_k = \left\{\mathcal{E}_\gamma(\mu_{k,m_k},\mu_k) \geq \frac{c}{\sqrt{m_k}} + \sqrt{\frac{ct_k}{m_k}}\right\}$$

for some sequence $t_1, t_2, \ldots$, and each $k \geq 1$. By Lemma 5.4.5,

$$\begin{aligned}\mathbb{P}(A) &\leq \delta,\\ \mathbb{P}(A_k) &= \mathbb{E}\mathbb{P}(A_k|X_1,\ldots,X_n) \leq 2\exp(-t_k).\end{aligned}$$

Taking $t_k = \log(k^2\pi^2/(3\delta))$, the union bound gives

$$\mathbb{P}\left(A \cup \bigcup_{k\geq 1} A_k\right) \leq 2\delta.$$

By the inequality $\mathcal{E}_\gamma(\mu_k,\nu) \leq \mathcal{E}_\gamma(\mu_k,\mu_{k,m_k}) + \mathcal{E}_\gamma(\mu_{k,m_k},\nu_n) + \mathcal{E}_\gamma(\nu_n,\nu)$ it follows that

$$\mathbb{P}\left(\exists k \,:\, \mathcal{E}_\gamma(\nu,\mu_k) > \mathcal{E}_\gamma(\mu_{k,m_k},\nu_n) + \frac{c}{\sqrt{n}} + \frac{c}{\sqrt{m_k}} + \sqrt{\frac{c\log(2/\delta)}{n}} + \sqrt{\frac{c\log(k^2\pi^2/(3\delta))}{m_k}}\right) \leq 2\delta.$$

Thus, by choosing $m_k \asymp n\log(k^2/\delta)/\log(1/\delta)$ we can conclude that there exists a constant c' depending only on β, d, γ such that

$$\mathbb{P}\left(\mathcal{E}_\gamma(\nu,\mu_k) \leq c'\sqrt{\frac{\log(1/\delta)}{n}} + \mathcal{E}_\gamma(\mu_{k,m_k},\nu_n), \forall k \geq 1\right) \geq 1 - 2\delta.$$

The final conclusion follows from Theorem 5.3.2. □

Note that our bound on the probability holds for all k simultaneously, which is made possible by the fact that m_k grows as $k \to \infty$. The empirical relevance of such a result is immediate: suppose we have proposed candidate generative models $\mu_1, \mu_2, \ldots$ (e.g. one after each period of training epochs, or from different training models) that is trained on an i.i.d. dataset $X_1, \ldots, X_n$ of size n from $\nu \in \mathcal{P}_S(\beta, d, C)$. A "verifier" only needs to request for m_k independent draws from the k'th candidate, and if we ever achieve $\mathcal{E}_{(\log n)^{-1}}(\mu_{k,m_k},\nu_n) \lesssim \sqrt{\log(1/\delta)/n}$ we can stop training and claim by Theorem 5.4.1 that we are a constant factor away from (near-)minimax optimality with probability $1-\delta$.

5.5 Suboptimality for Two-Sample Testing

So far in this paper we have shown how the empirical energy distance minimizer, while being mismatched with the target total variation loss, nevertheless achieves nearly minimax optimal

performance for density estimation tasks. Unfortunately, this surprising effect does not carry over to other statistical tasks, such as two-sample testing, which we describe in this section.

The task of two-sample testing over a family of distributions $\mathcal{P}$ is the following. Given two samples $(X, Y) \sim p^{\otimes n} \otimes q^{\otimes m}$ with unknown distribution, we need to distinguish between the hypotheses

$$H_0 : p = q \text{ and } p \in \mathcal{P}, \quad \text{versus} \quad H_1 : \mathsf{TV}(p, q) > \epsilon, \text{and } p, q \in \mathcal{P}$$

with vanishing type-I and type-II error. The special case of $m = \infty$ is known as *goodness-of-fit* testing, and for the class of smooth distributions it was famously solved by [110], who showed that in dimension $d = 1$ the problem is solvable with probability $1 - O(1)$ if and only if

$$n = \omega(\epsilon^{-\frac{2\beta+d/2}{\beta}}), \tag{5.5.1}$$

in which case a variant of the χ^2-test works. The case of general m, n and $d \geq 1$ was resolved in [10] who showed that the problem is solvable if and only if (5.5.1) holds with n replaced by $\min\{n, m\}$, using the very same χ^2-test; see also [139]. In the remainder of the section we focus on the $m = n$ case for simplicity.

In a recent paper [161], the following test statistic for two-sample testing was proposed:

$$T_{d,k}(p, q) = \max_{(w,b)\in\mathbb{S}^{d-1}\times[0,\infty)} \left| \mathbb{E}_{X\sim p} \left(w^\top X - b\right)_+^k - \mathbb{E}_{Y\sim q} \left(w^\top Y - b\right)_+^k \right|$$

where the arguments X, Y can be either discrete (e.g. via observed samples) or continuous densities. Note that here we take $(a)_+^0 = \mathbb{1}\{a \geq 0\}$ by convention. Specifically, the test proposed is to reject the null hypothesis when

$$T_{d,k}(p_n, q_n) \geq t_n, \tag{5.5.2}$$

where $p_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}, q_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ are empirical measures and the threshold that satisfies both $t_n = O(1)$ and $t_n = \omega(1/\sqrt{n})$. One of their main technical results [161, Theorem 6] asserts that the test (5.5.2) returns the correct hypothesis with probability $1 - O(1)$ asymptotically as $n \to \infty$ for any qualifying sequence $\{t_n\}_{n\geq 1}$ and fixed p, q. However, this result leaves open questions about the sample complexity of their test, and in particular, whether it is able to achieve known minimax rates. It turns out that our results imply that their test, at least in the $k = 0$ case, cannot attain the optimal two-sample testing sample complexity (5.5.1) over the smooth class $\mathcal{P}_S(\beta, d, C)$. To connect to our results, notice that

$$T_{d,0}(p, q) = \overline{d_H}(p, q)\,.$$

Proposition 5.5.1. *For all $d, \beta > 0$, there exists constants c, c' such that for all $\epsilon > 0$, there exists probability density functions p, q supported on the d-dimensional unit ball such that*

1. $\|p\|_{\beta,2}, \|q\|_{\beta,2} < c$,
2. $\|p - q\|_1 \asymp \|p - q\|_2 \asymp \epsilon$, *and*

3. the expected test statistic satisfies

$$\mathbb{E}[T_{d,0}(p_n, q_n)] \leq \frac{c'}{\sqrt{n}}$$

for any $n \leq (\log \frac{1}{\epsilon})^{-d} \epsilon^{-\frac{2\beta+d+1}{\beta}}$.

In other words, consistent testing using the statistic $T_{d,0}$ is impossible with $n = \tilde{O}(\epsilon^{-\frac{2\beta+d+1}{\beta}})$ samples, which is a far cry from the optimal sample complexity (5.5.1) attainable by the χ^2 test. The proof of Proposition 5.5.1 is given at Appendix D.5.

5.6 Conclusion

We analyzed the simple discriminating class of affine classifiers and proved its effectiveness in the ERM-GAN setting (5.1.2) within the Sobolev class $\mathcal{P}_S(\beta, d)$ and Gaussian mixtures $\mathcal{P}_G(d)$ with respect to the L^2 norm (see Theorem 5.4.1 and Corollary 5.4.2) and the total variation distance (see Theorem 5.1.1). Our findings affirm the rate's near-optimality for the considered classes of $\mathcal{P}_S$ and $\mathcal{P}_G$. Moreover, we present inequalities that interlink the $\mathcal{E}_\gamma$, TV, and L^2 distances, and demonstrate (in some cases) the tightness of these relationships via corresponding lower bound constructions (Appendix D.4). We also interpret the generalized energy distance in several ways that help advocate for its use in real applications. This work connects to a broader literature on the theoretical analysis of GAN-style models.

An interesting question emerges about the interaction between the expressiveness and concentration of the discriminator class. We found that the class of affine classifiers $\mathcal{D}_1$ is guaranteed to maintain some (potentially small) proportion of the total variation distance, and that it decays at the parametric rate between population and empirical distributions. Thus, we have traded off expressiveness for better concentration of the resulting IPM. As discussed in Section 5.1.2, Yatracos' estimator lies at the other end of this discriminator expressiveness-concentration trade-off: the distance d_Y is as expressive as total variation when restricted to the generator class $\mathcal{G}$, but $\sup_{\nu \in \mathcal{G}} \mathbb{E} d_Y(\nu, \nu_n)$ decays strictly slower than $1/\sqrt{n}$ for nonparametric classes $\mathcal{G}$. A downside compared to $\overline{d_H}$ is that (i) the Yatracos class $\mathcal{Y}$ requires knowledge of $\mathcal{G}$ while our $\mathcal{D}_1$ is oblivious to $\mathcal{G}$ and (ii) the distance d_Y is impractical to compute as it requires a covering of $\mathcal{G}$. Our question is: is it possible to find a class of sets $\mathcal{S} \subseteq 2^{\mathbb{B}(0,1)}$ that lies at an intermediate point on this trade-off? In other words, does $\mathcal{S}$ exist such that the ERM $\tilde{\nu}$ defined in (5.1.2) using the discriminator class $\mathcal{D} = \mathcal{S}$ is optimal over, say, $\mathcal{G} = \mathcal{P}_S$ and the induced distance converges slower than $1/\sqrt{n}$ but faster than $n^{-\beta/(2\beta+d)}$ between empirical and population measures? Would there be desiderata for a sample-efficient discriminator that has neither full expressiveness against total variation and does not concentrate at a parametric rate?

Appendix A

Appendix of "Likelihood-Free Hypothesis Testing"

A.1 Proof of achievability in Theorem 2.3.2 and 2.3.3

Let μ be a measure on the measurable space $(\mathcal{X}, \mathcal{F})$. Let $\{\phi_i\}_{i\in[r]}$ be a sequence of orthonormal functions in $L^2(\mu)$, where we use the notation $[r] =: \{1, 2, \ldots, r\}$. For $f \in L^2(\mu)$, define its projection onto the span of $\{\phi_1, \ldots, \phi_r\}$ as

$$P_r(f) =: \sum_{i\in[r]} \langle f\phi_i\rangle \phi_i,$$

where we write $\langle\cdot\rangle$ for integration with respect to μ and $\|\cdot\|_p$ for $\|\cdot\|_{L^p(\mu)}$. Given an i.i.d. sample $X = (X_1, \ldots, X_n)$ from some density f, define its empirical projection as

$$\widehat{P}_r[X] =: \sum_{i\in[r]} \left(\frac{1}{n}\sum_{j=1}^{n} \phi_i(X_j)\right) \phi_i.$$

We define the difference in L^2-distances statistics to be

$$T_{\mathsf{LF}} = \|\widehat{P}_r[X] - \widehat{P}_r[Z]\|_2^2 - \|\widehat{P}_r[Y] - \widehat{P}_r[Z]\|_2^2, \tag{A.1.1}$$

for an appropriate choice of μ and $\{\phi_j\}_{j\geq 1}$ depending on the class $\mathcal{P}$. Before calculating the mean and variance, we separate out the diagonal terms in T_{LF} thereby decomposing the statistic into two terms:

$$T_{\mathsf{LF}} =: T_{\mathsf{LF}}^{-\mathsf{d}} + \underbrace{\frac{1}{n^2}\sum_{i\in[r]}\sum_{j\in[n]} \left(\phi_i^2(X_j) - \phi_i^2(Y_j)\right)}_{=:D}, \tag{A.1.2}$$

which will simplify our proofs somewhat.

To ease notation in the results below, we define the quantities

$$\begin{aligned} A_{fgh} &= \langle f\left[P_r(g-h)\right]^2\rangle \\ B_{fg} &= \sum_{i=1}^{r} \langle f\phi_i P_r(g\phi_i)\rangle \end{aligned} \tag{A.1.3}$$

for $f, g, h \in L^2(\mu)$, assuming the quantities involved are well-defined. We are ready to state our meta-result from which we derive all our likelihood-free hypothesis testing upper bounds.

Proposition A.1.1. *Let f, g, h denote probability densities on $\mathcal{X}$ with respect to μ, and suppose we observe independent samples X, Y, Z of size n, n, m from f, g, h respectively. Then*

$$\begin{aligned}\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} &= \|P_r(f-h)\|_2^2 - \|P_r(g-h)\|_2^2 + \frac{1}{n}(\|P_r(g)\|_2^2 - \|P_r(f)\|_2^2)\\ \mathrm{var}(T_{\mathsf{LF}}^{-\mathsf{d}}) &\lesssim \frac{A_{ffh}+A_{ggh}}{n} + \frac{A_{hfg}}{m} + \frac{\|f+g+h\|_2^4 + |B_{fh}| + |B_{gh}|}{nm}\\ &\quad + \frac{|B_{ff}| + |B_{gg}| + \|f+g+h\|_2^4 + \sqrt{A_{ff0}A_{ffh} + A_{gg0}A_{ggh}}}{n^2}\\ &\quad + \frac{|B_{ff}| + |B_{gg}| + \|f+g+h\|_2^4 + A_{ff0} + A_{gg0}}{n^3},\end{aligned}$$

where the implied constant is universal.

Proposition A.1.1 is used to test (LF) by rejecting the null whenever $T_{\mathsf{LF}}^{-\mathsf{d}} \geq 0$. To prove that this procedure performs well we show that $T_{\mathsf{LF}}^{-\mathsf{d}}$ concentrates around its mean by Chebyshev's inequality. For this we find sufficient conditions on the sample sizes n, m so that $(\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}})^2 \gtrsim \mathrm{var}(T_{\mathsf{LF}}^{-\mathsf{d}})$ for a small enough implied constant on the left.

While Proposition A.1.1 is enough to conclude the proof of our main theorems, notice that it uses the statistic $T_{\mathsf{LF}}^{-\mathsf{d}}$ which has the diagonal terms removed. For completeness we show that rejecting when $T_{\mathsf{LF}} \geq 0$ is also minimax optimal, that is, the diagonal term D in (A.1.2) can be included without degrading performance.

A.1.1 The class $\mathcal{P}_{\mathsf{Db}}$

Proposition A.1.2. *For any $C > 1$ there exists a constant $c > 0$ such that*

$$\mathcal{R}_{\mathsf{rLF}}(\epsilon, \mathcal{P}_{\mathsf{Db}}(k, C), \mathsf{B}_\cdot) \supset c\left\{m \geq 1/\epsilon^2, n \geq \sqrt{k}/\epsilon^2, mn \geq k/\epsilon^4\right\},$$

where $\mathsf{B}_u = \{u \in \mathcal{P}_{\mathsf{Db}}(k, C) : \|u - v\|_2 \leq \epsilon/(2\sqrt{k})\}$.

Proof. **Choice of μ and ϕ.** Take $\mathcal{X} = [k]$ and let $\mu = \sum_{i=1}^k \delta_i$ be the counting measure. Let $\phi_i(j) = \mathbb{1}_{\{i=j\}}$ and choose $r = k$ so that $P_r = P_k$ is the identity. By the Cauchy-Schwarz inequality $\|u\|_1 \leq \sqrt{k}\|u\|_2$ for all $u \in \mathbb{R}^k$.

Applying Proposition A.1.1. Recall the notation of Proposition A.1.1, so that f, g, h are the pmfs of $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Z}}$ respectively. We analyse the performance of the test $\mathbb{1}\{T_{\mathsf{LF}}^{-\mathsf{d}} \geq 0\}$ under the null hypothesis, the proof under the alternative is analogous. The inequality

$$\|f-h\|_2 \leq \frac{\epsilon}{2\sqrt{k}} \leq \frac{\|f-g\|_1}{4\sqrt{k}} \leq \frac{\|f-g\|_2}{4}$$

along with the reverse triangle inequality yields

$$\begin{aligned}\|g-h\|_2^2 - \|f-h\|_2^2 &\geq (\|f-g\|_2 - \|f-h\|_2)^2 - \|f-h\|_2^2\\ &= \|f-g\|_2^2 - 2\|f-g\|_2\|f-h\|_2\\ &\geq \|f-g\|_2^2/2.\end{aligned}$$

Combining the above inequality with Proposition A.1.1, we get that $-\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} \geq \|f-g\|_2^2/2+R$, where the residual term R can be bounded as

$$\begin{aligned}|R| &= \left|\frac{\|f\|_2^2 - \|g\|_2^2}{n}\right| \\ &\leq 2C\frac{\|f-g\|_2}{n\sqrt{k}}.\end{aligned}$$

Therefore, $-\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} \geq \frac{\|f-g\|_2^2}{4}$ holds provided $2C\|f-g\|_2/(n\sqrt{k}) \leq \|f-g\|_2^2/4$, which in turn is implied by $n \gtrsim 1/\epsilon$ and is thus always satisfied.

Turning towards the variance, we apply Proposition A.1.1 to see that

$$\operatorname{var}(T_{\mathsf{LF}}^{-\mathsf{d}}) \lesssim \frac{\|f-g\|_2^2}{k}\left(\frac{1}{n}+\frac{1}{m}\right) + \frac{1}{k}\left(\frac{1}{n^2}+\frac{1}{nm}\right), \tag{A.1.4}$$

where we use the trivial bounds

$$\begin{aligned}\|f+g+h\|_2 &\lesssim \sqrt{\frac{C}{k}} \lesssim \sqrt{\frac{1}{k}} \\ |B_{ff}| + |B_{gg}| + |B_{fh}| + |B_{gh}| &\lesssim \frac{C}{k} \lesssim \frac{1}{k} \\ A_{ffh} + A_{ggh} + A_{hfg} &\lesssim \frac{C}{k}\|f-g\|_2^2 \lesssim \frac{1}{k}\|f-g\|_2^2 \\ A_{ff0} + A_{gg0} &\lesssim \left(\frac{C}{k}\right)^2 \lesssim \frac{1}{k^2}.\end{aligned}$$

Applying Chebyshev's inequality and looking at each term separately in (A.1.4) and using that $\|f-g\|_2 \geq \epsilon/(2\sqrt{k})$ yields the desired bounds on n, m.

The diagonal. While the above test using $T_{\mathsf{LF}}^{-\mathsf{d}}$ already achieves the minimax optimal sample complexity, here we show for completeness that the diagonal D, defined in (A.1.2), can be included without degrading the test's performance. Indeed, we always have

$$\begin{aligned}D &= \frac{1}{n^2}\sum_{i\in[r]}\sum_{j\in[n]}\left(\mathbb{1}\{X_j=i\}^2 - \mathbb{1}\{Y_j=i\}^2\right) \\ &= 0.\end{aligned}$$

Therefore, trivially, the test $\mathbb{1}\{T_{\mathsf{LF}} \geq 0\}$ has the same performance as the one analyzed above. □

A.1.2 The class $\mathcal{P}_{\mathsf{H}}$

Proposition A.1.3. *For every $C > 1, \beta > 0$ and $d \geq 1$ there exist two constants $c, c_{\mathsf{r}} > 0$ such that*

$$\mathcal{R}_{\mathsf{rLF}}(\epsilon, \mathcal{P}_{\mathsf{H}}(\beta,d,C), \mathsf{B}_\cdot) \supset c\left\{m \geq 1/\epsilon^2, n \geq 1/\epsilon^{(2\beta+d/2)/\beta}, mn \geq 1/\epsilon^{2(2\beta+d/2)/\beta}\right\}, \tag{A.1.5}$$

where $\mathsf{B}_u = \{v \in \mathcal{P}_{\mathsf{H}}(\beta,d,C) : \|v-u\|_2 \leq c_{\mathsf{r}}\epsilon\}$.

Proof. **Choice of μ and ϕ.** Take $\mathcal{X} = [0,1]^d$, let μ the Lebesgue measure on $\mathcal{X}$. Let $\{\phi_i\}_{1\le i\le \kappa^d}$ be the indicators of the cells of the regular grid with κ^d bins, normalized to have $L^2(\mu)$-norm equal to 1, that is, the indicator is multiplied by κ^d, which is one over the volume of one grid cell. By [10, Lemma 7.2] for any resolution $r = \kappa^d$ and $u \in \mathcal{C}(\beta, d, 2C)$ we have

$$\|P_r(u)\|_2 \ge c_1\|u\|_2 - c_2\kappa^{-\beta} \tag{A.1.6}$$

for constants $c_1, c_2 > 0$ that don't depend on r. In particular, the inequalities

$$\|P_r(u)\|_2 \ge \frac{c_1}{2}\|u\|_2 \tag{A.1.7}$$

holds for any $\|u\|_2 \ge \epsilon$ provided we choose $\kappa = \left(\frac{2c_2}{c_1\epsilon}\right)^{1/\beta}$.

Applying Proposition A.1.1. Recall the notation of Proposition A.1.1 so that f, g, h are the μ-densities of $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Z}}$. We analyse the performance of the test $\mathbb{1}\{T_{\mathsf{LF}}^{-\mathsf{d}} \ge 0\}$ under the null hypothesis, the proof under the alternative is analogous. Let the radius of robustness be $c_{\mathsf{r}} = c_1/4$, and set $\kappa = \left(\frac{2c_2}{c_1\epsilon}\right)^{1/\beta}$. Then we have

$$\|P_r(f-h)\|_2 \le c_{\mathsf{r}}\epsilon = \frac{c_{\mathsf{r}}}{2}\|f-g\|_2 \le \frac{c_{\mathsf{r}}}{c_1}\|P_r(f-g)\|_2$$

by taking $u = f - g$ in (A.1.7). Using the reverse triangle inequality we obtain

$$\begin{aligned}
\|P_r(g-h)\|_2^2 - \|P_r(f-h)\|_2^2 &\ge \left(\|P_r(f-g)\|_2 - \|P_r(f-h)\|_2\right)^2 - \|P_r(f-h)\|_2^2 \\
&= \|P_r(f-g)\|_2^2 - 2\|P_r(f-g)\|_2\|P_r(f-h)\|_2 \\
&\ge \|P_r(f-g)\|_2^2(1 - 2\frac{c_{\mathsf{r}}}{c_1}) \\
&= \|P_r(f-g)\|_2^2/2
\end{aligned}$$

Combining the above inequality with Proposition A.1.1, we see that $-\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} \ge \|P_r(f-g)\|_2^2/2 + R$ where the residual term R can be bounded as

$$\begin{aligned}
|R| &= \left|\frac{\|f\|_2^2 - \|g\|_2^2}{n}\right| \\
&\le 2C\frac{\|f-g\|_2}{n}.
\end{aligned}$$

Therefore, the inequality $-\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} \ge \|P_r(f-g)\|_2^2/4$ holds provided $2C\|f-g\|_2/n \le \|P_r(f-g)\|_2^2/4$, which in turn is implied by $n \gtrsim 1/\epsilon$ and is thus always satisfied.

Turning to the variance, using Proposition A.1.1 we obtain

$$\mathrm{var}(T_{\mathsf{LF}}^{-\mathsf{d}}) \lesssim \|P_r(f-g)\|_2^2\left(\frac{1}{n} + \frac{1}{m}\right) + \epsilon^{-d/\beta}\left(\frac{1}{n^2} + \frac{1}{nm}\right), \tag{A.1.8}$$

where we apply the trivial inequalities

$$\begin{aligned}
\|f+g+h\|_2 &\lesssim \sqrt{C} \lesssim 1 \\
|B_{ff}| + |B_{gg}| + |B_{fh}| + |B_{gh}| &\lesssim Cr = C\kappa^d \asymp \epsilon^{-d/\beta} \\
A_{ffh} + A_{ggh} + A_{hfg} &\lesssim C\|P_r(f-g)\|_2^2 \lesssim \|P_r(f-g)\|_2^2 \\
A_{ff0} + A_{gg0} &\lesssim C^2 \lesssim 1.
\end{aligned}$$

Applying Chebyshev's inequality and looking at each term separately in (A.1.8) and using that $\|P_r(f-g)\|_2 \gtrsim \|f-g\|_2 \geq \|f-g\|_1 \geq 2\epsilon$ yields the desired bounds on n, m.

The diagonal. While the above test using $T_{\mathsf{LF}}^{-\mathsf{d}}$ already achieves the minimax optimal sample complexity, for completeness we also note that including the diagonal terms D defined in (A.1.2) doesn't degrade performance. This follows from the simple fact that $D = 0$, which is true for reasons analogous to the case of $\mathcal{P}_{\mathsf{Db}}$ that we already covered. □

A.1.3 The class $\mathcal{P}_{\mathsf{G}}$

Proposition A.1.4. *For all $s, C > 0$ there exists a constant $c > 0$ such that*

$$\mathcal{R}_{\mathsf{rLF}}(\epsilon, \mathcal{P}_{\mathsf{G}}(s,C), \mathsf{B}.) \supset c\left\{m \geq 1/\epsilon^2, n \geq 1/\epsilon^{(2s+1/2)/s}, mn \geq 1/\epsilon^{2(2s+1/2)/s}\right\},$$

where $\mathsf{B}_{\mu_\theta} = \{\mu_{\theta'} : \theta' \in \mathcal{E}(s,C), \|\theta - \theta'\|_2 \leq \epsilon/4\}$.

Proof. **Choosing μ and ϕ.** Let $\mathcal{X} = \mathbb{R}^{\mathbb{N}}$ be the set of infinite sequences and take as the base measure $\mu = \otimes_{d=1}^{\infty}\mathcal{N}(0,1)$, the infinite dimensional standard Gaussian. For $\theta \in \ell^2$ write $\mu_\theta = \otimes_{d=1}^{\infty}\mathcal{N}(\theta_i, 1)$ so that $\mu_0 = \mu$. Take the orthonormal functions $\phi_i(x) = x_i$ in $L^2(\mu)$ for $i \geq 1$, so that

$$P_r\left(\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu}\right) = \sum_{i=1}^{r} x_i\theta_i.$$

Let $\theta, \theta' \in \mathcal{E}(s,C)$ with $\mathsf{TV}(\mu_\theta, \mu_{\theta'}) \geq \epsilon$. By direct computation we obtain

$$\|P_r\left(\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu} - \frac{\mathrm{d}\mu_{\theta'}}{\mathrm{d}\mu}\right)\|_2^2 = \sum_{i=1}^{r}(\theta_i - \theta_i')^2 \geq \|\theta-\theta'\|_2^2 - r^{-2s}\sum_{i>r}(\theta_i-\theta_i')^2 i^{2s} \geq \|\theta-\theta'\|_2^2 - 4C^2 r^{-2s}. \tag{A.1.9}$$

In particular, the inequality

$$\|P_r\left(\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu} - \frac{\mathrm{d}\mu_{\theta'}}{\mathrm{d}\mu}\right)\|_2^2 \geq \frac{1}{2}\|\theta-\theta'\|_2^2 \tag{A.1.10}$$

holds for all $\theta, \theta' \in \mathcal{E}(s,C)$ with $\|\theta - \theta'\|_2 \geq \epsilon$, provided we take $r = (4C/\epsilon)^{1/s}$.

Applying Proposition A.1.1. Recall the notation of Proposition A.1.1, and let f, g, h be the μ-densities of $\mathbb{P}_{\mathsf{X}} = \mu_{\theta_{\mathsf{X}}}, \mathbb{P}_{\mathsf{Y}} = \mu_{\theta_{\mathsf{Y}}}, \mathbb{P}_{\mathsf{Z}} = \mu_{\theta_{\mathsf{Z}}}$ respectively. We analyse the test $\mathbb{1}\{T_{\mathsf{LF}}^{-\mathsf{d}} \geq 0\}$ only under the null hypothesis, as the analysis under the alternative is analogous. Note also that by Lemma A.2.2 the inequality

$$\mathsf{TV}(\mu_\theta, \mu_{\theta'}) \leq \mathsf{H}(\mu_\theta, \mu_{\theta'}) = \sqrt{2(1-\exp(-\|\theta-\theta'\|_2^2/8))} \leq \frac{\|\theta-\theta'\|_2}{2}$$

holds for any $\theta, \theta' \in \ell^2$. Therefore, we have

$$\|P_r(f-h)\|_2 \leq \frac{\epsilon}{4} \leq \frac{\mathsf{TV}(\mu_\theta, \mu_{\theta'})}{4} \leq \frac{\|\theta-\theta'\|_2}{8} \leq \frac{\|P_r(f-g)\|_2}{4}$$

by (A.1.10).

By the reverse triangle inequality we have

$$\begin{aligned}\|P_r(g-h)\|_2^2 - \|P_r(f-h)\|_2^2 &\geq \left(\|P_r(f-g)\|_2 - \|P_r(f-h)\|_2\right)^2 - \|P_r(f-h)\|_2^2 \\ &= \|P_r(f-g)\|_2^2 - 2\|P_r(f-g)\|_2\|P_r(f-h)\|_2 \\ &\geq \|P_r(f-g)\|_2^2/2\end{aligned}$$

Combining the inequality above with Proposition A.1.1, we see that $-\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} \geq \|P_r(f-g)\|_2^2/2 + R$, where the residual term R can be bounded as

$$\begin{aligned}|R| &= \left|\frac{\|P_r(f)\|_2^2 - \|P_r(g)\|_2^2}{n}\right| \\ &\leq 2C\frac{\|P_r(f-g)\|_2}{n}.\end{aligned}$$

Therefore, $-\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} \geq \|P_r(f-g)\|_2^2/4$ provided $2C\|P_r(f-g)\|_2/n \leq \|P_r(f-g)\|_2^2/4$, which in turn is implied by $n \gtrsim 1/\epsilon$ and is therefore always satisfied.

Let us turn to the variance of the statistic. Let u, v, t be the μ-densities of the distributions $\mu_\theta, \mu_{\theta'}, \mu_{\theta''}$ for some vectors $\theta, \theta', \theta'' \in \mathcal{E}(s, C)$ in the Sobolev ellipsoid. Straightforward calculations involving Gaussian random variables produce

$$\begin{aligned}A_{uvt} &= \sum_{ij}^{r}(\mathbb{1}(i=j) + \theta_i\theta_j)(\theta'_i - \theta''_i)(\theta'_j - \theta''_j) \leq (1+C^2)\|P_r(v-t)\|_2^2 \\ &\lesssim \|P_r(v-t)\|_2^2 \lesssim C^2 \lesssim 1 \\ \|u\|_2 &= \exp\left(\frac{1}{2}\|\theta\|_2^2\right) \leq \exp(C^2/2) \lesssim 1 \\ B_{uv} &= \sum_{i=1}^{r}\left(1 + \theta_i^2 + {\theta'}_i^2 + \theta_i\theta'_i\sum_{j=1}^{r}\theta_j{\theta'}_j\right) \\ &\leq r + 2C^2 + C^4 \\ &\lesssim r.\end{aligned}$$

Applying Proposition A.1.1 tells us that

$$\mathrm{var}(T_{\mathsf{LF}}^{-\mathsf{d}}) \lesssim \|P_r(f-g)\|_2^2\left(\frac{1}{n} + \frac{1}{m}\right) + \epsilon^{-1/s}\left(\frac{1}{n^2} + \frac{1}{nm}\right) \tag{A.1.11}$$

Applying Chebyshev's inequality and looking at each term separately in (A.1.11) and using that $\mathsf{TV}(\mu_\theta, \mu_{\theta'}) \lesssim \|P_r(f-g)\|$ yields the desired bounds on n, m.

The diagonal. While the above test using $T_{\mathsf{LF}}^{-\mathsf{d}}$ already achieves the minimax optimal sample complexity, for completeness we show that including the diagonal terms D defined in

(A.1.2) doesn't degrade performance. To this end we compute

$$\begin{aligned}\mathbb{E}D &= \mathbb{E}\frac{1}{n^2}\sum_{i\in[r]}\sum_{j\in[n]}\left(\phi_i^2(X_j)-\phi_i^2(Y_j)\right)\\ &= \frac{1}{n}\sum_{i\in[r]}(\theta_{\mathsf{X},i}^2-\theta_{\mathsf{Y},i}^2)\\ &\le \frac{1}{n}\|\theta_{\mathsf{X}}+\theta_{\mathsf{Y}}\|_2\sqrt{\sum_{i\in[r]}(\theta_{\mathsf{X},i}-\theta_{\mathsf{Y},i}{}^2)}\\ &\le 2C\frac{\|P_r(f-g)\|_2}{n}.\end{aligned}$$

We see that $|\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}}| \gtrsim |\mathbb{E}D|$ as soon as $n \gtrsim 1/\epsilon$. Turning to the variance, we have

$$\begin{aligned}\mathrm{var}(D) &= \frac{1}{n^3}\sum_{i\in[r]}\left(\mathrm{var}(\phi_i^2(X_1))+\mathrm{var}(\phi_i^2(Y_1))\right)\\ &\lesssim \frac{rC^2}{n^3},\end{aligned}$$

and so the diagonal terms do not inflate the variance by more than a constant factor. Therefore, the sample complexity of the test is unchanged. □

A.1.4 The class $\mathcal{P}_{\mathsf{D}}$

Proposition A.1.5. *Let $\alpha = 1 \vee \left(\frac{k}{n}\wedge\frac{k}{m}\right)$. There exist constants $c, c', c_{\mathsf{r}} > 0$ such that*

$$\mathcal{R}_{\mathsf{rLF}}(\epsilon, \mathcal{P}_{\mathsf{D}}(k), \mathsf{B}_\cdot) \supset \frac{c}{\log(k)}\left\{m \ge 1/\epsilon^2, n \ge \sqrt{k\alpha}/\epsilon^2, mn \ge k\alpha/\epsilon^4\right\},$$

where $\mathsf{B}_u = \{v : \|u-v\|_2 \le c_{\mathsf{r}}\epsilon/\sqrt{k}, \|v/u\|_\infty \le c'\}$.

Proof. **Choosing μ and ϕ.** As for $\mathcal{P}_{\mathsf{Db}}$, we take $\mathcal{X} = [k]$, $\mu = \sum_{i=1}^k \delta_i$, $\phi_i(j) = \mathbb{1}_{\{i=j\}}$ and $r = k$. By the Cauchy-Schwarz inequality $\|h\|_1 \le \sqrt{k}\|h\|_2$ for all $h \in \mathbb{R}^k$.

Reducing to the small-norm case. Before applying Proposition A.1.1 we need to 'pre-process' our distributions. For an in-depth explanation of this technique see [64, 80]. Recall that we write f, g, h for the probability mass functions of $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Z}}$ respectively, from which we observe the samples X, Y, Z of size n, n, m respectively. Recall also that the null hypothesis is that $\|f-h\|_2 \le c_{\mathsf{r}}\epsilon/\sqrt{k}$ while the alternative says that $\|g-h\|_2 \le c_{\mathsf{r}}\epsilon/\sqrt{k}$, with $\|f-g\|_2 \ge 2\epsilon/\sqrt{k}$ guaranteed under both. In the following section we use the standard inequality $\mathbb{P}(\lambda - x \ge \mathrm{Poi}(\lambda)) \le \exp(-\frac{x^2}{2(\lambda+x)})$ valid for all $x \ge 0$ repeatedly. We also utilize the identity

$$\mathbb{E}\left[\frac{1}{\mathrm{Poi}(\lambda)+1}\right] = \begin{cases}1 & \text{if } \lambda = 0\\ \frac{1-e^{-\lambda}}{\lambda} & \text{if } \lambda > 0,\end{cases} \tag{A.1.12}$$

which is easily verified by direct calculation. Finally, the following Lemma will come handy.

Proposition A.1.6. *[80, Corollary 11.6] Given t samples from an unknown discrete distribution p, there exists an algorithm that produces an estimate $\widehat{\|p\|_2^2}$ with the property*

$$\mathbb{P}(\widehat{\|p\|_2^2} \notin (\frac{1}{2}\|p\|_2^2, \frac{3}{2}\|p\|_2^2)) \lesssim \frac{1}{\|p\|_2 t},$$

where the implied constant is universal.

First we describe a random "filter" $F : \mathcal{P}_{\mathsf{D}}(k) \to \mathcal{P}_{\mathsf{D}}(K)$ that maps distributions on $[k]$ to distributions on the inflated alphabet $[K]$. Let $(n_{\mathsf{X}}, n_{\mathsf{Y}}, n_{\mathsf{Z}}) = \frac{1}{2}(n \wedge k, n \wedge k, m \wedge k)$ and let $N^{\mathsf{X}} \sim \mathrm{Poi}(n_{\mathsf{X}}/2)$ independently of all other randomness, and define $N^{\mathsf{Y}}, N^{\mathsf{Z}}$ similarly. We take the first $N^{\mathsf{X}}, N^{\mathsf{Y}}, N^{\mathsf{Z}}$ samples from the data sets X, Y, Z respectively. In the event $N^{\mathsf{X}} \vee N^{\mathsf{Y}} > n$ or $N^{\mathsf{Z}} > m$ let our output to the likelihood-free hypothesis test be arbitrary, this happens with exponentially small probability. Let N_i^{X} be the number of the samples $X_1, \ldots, X_{N^{\mathsf{X}}}$ falling in bin i, so that $N_i^{\mathsf{X}} \sim \mathrm{Poi}(n_{\mathsf{X}} f_i/2)$ independently for each $i \in [k]$, and define $N_i^{\mathsf{Y}}, N_i^{\mathsf{Z}}$ analogously. The filter F is defined as follows:

> divide each support element $i \in \{1, 2, \ldots, k\}$ uniformly into $1 + N_i^{\mathsf{X}} + N_i^{\mathsf{Y}} + N_i^{\mathsf{Z}}$ bins.

The filter has the following properties trivially:

1. The construction succeeds with probability $\geq 1 - 3\exp(-n \wedge m \wedge k/16)$, focus on this event from here on.
2. The construction uses at most $n_{\mathsf{X}}, n_{\mathsf{Y}}, n_{\mathsf{Z}}$ samples from X, Y, Z respectively and satisfies $K \leq 5k/2$.
3. For any $u, v \in \mathcal{P}_{\mathsf{D}}(k)$ we have $\mathsf{TV}(F(u), F(v)) = \mathsf{TV}(u, v)$ and $\|F(u) - F(v)\|_2 \leq \|u - v\|_2$.
4. Given a sample from an unknown $u \in \mathcal{P}_{\mathsf{D}}(k)$ we can generate a sample from $F(u)$ and vice-versa.

Let $\tilde{f} =: F(f)$ be the probability mass function after processing and define $\tilde{g}, \tilde{h}$ analogously. By properties $1-2$ of the filter, we may assume with probability 99% that the new alphabet's size is at most $5k/2$ and that we used at most half of our samples X, Y, Z. We immediately get $2\epsilon \leq \|f - g\|_1 = \|\tilde{f} - \tilde{g}\|_1 \leq \sqrt{5k/2}\|\tilde{f} - \tilde{g}\|_2$ and $\|\tilde{f} - \tilde{h}\|_2 \leq \|f - h\|_2, \|\tilde{g} - \tilde{h}\|_2 \leq \|g - h\|_2$. Notice that

$$\sum_{i \in [K]} \tilde{f}_i \tilde{g}_i = \sum_{i \in [k]} \frac{f_i g_i}{1 + N_i^{\mathsf{X}} + N_i^{\mathsf{Y}} + N_i^{\mathsf{Z}}}$$

holds, and similar statements can be derived for the inner product between $\tilde{f}, \tilde{h}$ etc. Recall that we set

$$\alpha = \max\left\{1, \min\left\{\frac{k}{n}, \frac{k}{m}\right\}\right\}.$$

Adopting the convention $0/0 = 1$ and using (A.1.12) we can bound inner products between the mass functions as

$$\mathbb{E}\left[B_{\tilde{f}\tilde{h}} + B_{\tilde{g}\tilde{h}}\right] = \mathbb{E}\left[\langle \tilde{f}\tilde{h}\rangle + \langle \tilde{g}\tilde{h}\rangle\right] \le 4\sum_{i\in[k]} \frac{f_i h_i + g_i h_i}{(n\wedge k)(f_i+g_i) + (m\wedge k)h_i} \le \frac{8}{(n\vee m)\wedge k} = \frac{8\alpha}{k}$$

$$\mathbb{E}\left[B_{\tilde{f}\tilde{f}} + B_{\tilde{g}\tilde{g}}\right] = \mathbb{E}\left[\|\tilde{f}\|_2^2 + \|\tilde{g}\|_2^2\right] \le 4\sum_{i\in[k]} \frac{f_i^2 + g_i^2}{(n\wedge k)(f_i+g_i) + (m\wedge k)h_i} \le \frac{8}{n\wedge k}$$

$$\mathbb{E}\|\tilde{h}\|_2^2 \le 4\sum_{i\in[k]} \frac{h_i^2}{(n\wedge k)(f_i+g_i) + (m\wedge k)h_i} \le \frac{4}{m\wedge k}.$$

By Markov's inequality we may assume that the inequalities in the preceding display hold not only in expectation but with 99% probability overall with universal constants. Notice that under the null hypothesis $\|\tilde{f} - \tilde{h}\|_2 \le c_{\mathsf{r}}\epsilon/\sqrt{k}$ and thus $\|\tilde{f}\|_2 \le \|\tilde{h}\|_2 + c_{\mathsf{r}}\epsilon/\sqrt{k} \le \|\tilde{f}\|_2 + 2c_{\mathsf{r}}\epsilon/\sqrt{k}$, and similarly with $\tilde{f}$ replaced by $\tilde{g}$ under the alternative. We restrict our attention to $c_{\mathsf{r}} \in (0,1)$ so that c_{r} is treated as a constant where appropriate. Notice that $\epsilon/\sqrt{k} \lesssim 1/\sqrt{(n\vee m)\wedge k}$ holds trivially. Thus, we obtain $\|\tilde{f}\|_2 \vee \|\tilde{h}\|_2 \le c/\sqrt{(m\vee n)\wedge k}$ under the null and $\|\tilde{g}\|_2 \vee \|\tilde{h}\|_2 \le c/\sqrt{(n\vee m)\wedge k}$ under the alternative for a universal constant c. We would like to ensure that

$$\|\tilde{f}\|_2 \vee \|\tilde{g}\|_2 \vee \|\tilde{h}\|_2 \lesssim \frac{1}{\sqrt{(m\vee n)\wedge k}} = \sqrt{\frac{\alpha}{k}}. \tag{A.1.13}$$

To this end we apply Proposition A.1.6 using $(n/4, n/4)$ of the remaining, transformed but otherwise untouched X, Y samples. Let $\widehat{\|\tilde{f}\|_2^2}, \widehat{\|\tilde{g}\|_2^2}$ denote the estimates, which lie in $(\frac{1}{2}\|\tilde{f}\|_2^2, \frac{3}{2}\|\tilde{f}\|_2^2)$ and $(\frac{1}{2}\|\tilde{g}\|_2^2, \frac{3}{2}\|\tilde{g}\|_2^2)$ respectively, with probability at least $1 - \mathcal{O}((\|\tilde{f}\|_2^{-1} + \|\tilde{g}\|_2^{-1})/n) \ge 1 - \mathcal{O}(\sqrt{k}/n)$, since $\|\tilde{f}\|_2 \wedge \|\tilde{g}\|_2 \ge \sqrt{2/(5k)}$ by the Cauchy-Schwarz inequality. Assuming that $n \gtrsim \sqrt{k}$ this probability can be taken to be arbitrarily high, say 99%. Now we perform the following procedure: if $\widehat{\|\tilde{f}\|_2^2} > \frac{3}{2}c^2/((n\vee m)\wedge k)$ reject the null hypothesis, otherwise if $\widehat{\|\tilde{g}\|_2^2} > \frac{3}{2}c^2/((n\vee m)\wedge k)$ accept the null hypothesis, otherwise proceed with the assumption that (A.1.13) holds. By design this process, on our 97% $\le$ probability event of interest, correctly identifies the hypothesis or correctly concludes that (A.1.13) holds. The last step of the reduction is ensuring that the quantities $A_{\tilde{f}\tilde{f}\tilde{h}}, A_{\tilde{g}\tilde{g}\tilde{h}}, A_{\tilde{h}\tilde{f}\tilde{g}}, A_{\tilde{f}\tilde{f}0}, A_{\tilde{g}\tilde{g}0}$ are small. The first two and last two may be bounded easily as

$$\begin{aligned}
A_{\tilde{f}\tilde{f}\tilde{h}} + A_{\tilde{g}\tilde{g}\tilde{h}} &= \langle \tilde{f}(\tilde{f}-\tilde{h})^2\rangle + \langle \tilde{g}(\tilde{g}-\tilde{h})^2\rangle \\
&\le \|\tilde{f}\|_2\|\tilde{f}-\tilde{h}\|_4^2 + \|\tilde{g}\|_2\|\tilde{g}-\tilde{h}\|_4^2 \\
&\lesssim \frac{\|\tilde{f}-\tilde{h}\|_2^2 + \|\tilde{g}-\tilde{h}\|_2^2}{\sqrt{(n\vee m)\wedge k}} \\
&\lesssim \frac{\|\tilde{f}-\tilde{g}\|_2^2 + c_{\mathsf{r}}^2\epsilon^2/k}{\sqrt{(n\vee m)\wedge k}} \lesssim \frac{\|\tilde{f}-\tilde{g}\|_2^2}{\sqrt{(n\vee m)\wedge k}} = \sqrt{\frac{\alpha}{k}}\|\tilde{f}-\tilde{g}\|_2^2 \\
A_{\tilde{f}\tilde{f}0} + A_{\tilde{g}\tilde{g}0} &= \|\tilde{f}\|_3^3 + \|\tilde{g}\|_3^3 \le \|\tilde{f}\|_2^3 + \|\tilde{g}\|_2^3 \lesssim \frac{1}{((n\vee m)\wedge k)^{3/2}} = \left(\frac{\alpha}{k}\right)^{3/2}.
\end{aligned} \tag{A.1.14}$$

To bound $A_{\tilde{h}\tilde{f}\tilde{g}}$ we need a more sophisticated method. Recall that by definition

$$A_{\tilde{h}\tilde{f}\tilde{g}} = \sum_{i\in[k]} \frac{h_i(f_i - g_i)^2}{(1 + N_i^{\mathsf{X}} + N_i^{\mathsf{Y}} + N_i^{\mathsf{Z}})^2}.$$

Fix an $i \in [k]$ and let $P =: N_i^{\mathsf{X}} + N_i^{\mathsf{Y}} + N_i^{\mathsf{Z}} \sim \mathrm{Poi}((n \wedge k)(f_i + g_i)/4 + (m \wedge k)h_i/4)$ and take a constant $c > 0$ to be specified. We have

$$\mathbb{P}\left(\frac{1}{1+P} > c\log(k)\frac{1}{\mathbb{E}P}\right) = \begin{cases} 0 & \text{if } \mathbb{E}P \le c\log(k) \\ \mathbb{P}\left(\mathbb{E}P - \left(\mathbb{E}P\left(1 - \frac{1}{c\log(k)}\right) + 1\right) > P\right) & \text{if } \mathbb{E}P > c\log(k). \end{cases}$$

Assuming that i is such that $\mathbb{E}P \ge c\log(k)$ and taking k large enough so that $c\log(k) \ge 2$, we can proceed as

$$\begin{aligned} \mathbb{P}\left(\mathbb{E}P - \left(\mathbb{E}P\left(1 - \frac{1}{c\log(k)}\right) + 1\right) > P\right) &\le \exp\left(-\frac{1}{2}\frac{(\mathbb{E}P(1 - \frac{1}{c\log(k)}) + 1)^2}{\mathbb{E}P(2 - \frac{1}{c\log(k)}) + 1}\right) \\ &\le \exp\left(-\frac{1}{16}\mathbb{E}P\right) \\ &\le \frac{1}{k^{c/16}}. \end{aligned}$$

Choosing $c = 32$ and taking a union bound, the inequality

$$A_{\tilde{h}\tilde{f}\tilde{g}} \lesssim \frac{\log(k)}{m \wedge k} \sum_{i\in[k]} \frac{(f_i - g_i)^2}{1 + N_i^{\mathsf{X}} + N_i^{\mathsf{Y}} + N_i^{\mathsf{Z}}} \asymp \frac{\log(k)}{m \wedge k}\|\tilde{f} - \tilde{g}\|_2^2$$

holds with probability at least $1 - 1/k$. Using that $\|h/f\|_\infty \wedge \|h/g\|_\infty \lesssim 1$ by assumption, we obtain $A_{\tilde{h}\tilde{f}\tilde{g}} \lesssim \frac{\log(k)}{n\wedge k}\|\tilde{f} - \tilde{g}\|_2^2$ similarly. Combining the two bounds yields

$$A_{\tilde{h}\tilde{f}\tilde{g}} \lesssim \frac{\log(k)}{(m \vee n) \wedge k}\|\tilde{f} - \tilde{g}\|_2^2 = \frac{\log(k)\alpha}{k}\|\tilde{f} - \tilde{g}\|_2^2. \tag{A.1.15}$$

To summarize, under the assumptions that $n \gtrsim \sqrt{k}$, and at the cost of inflating the alphabet size to at most $\frac{5}{2}k$ and a probability of error at most $3\% + \frac{1}{k}$, we may assume that the inequalities (A.1.13), (A.1.14) and (A.1.15) hold with universal constants.

Applying Proposition A.1.1. We only analyse the type-I error, as the type-II error follows analogously. As explained earlier, we apply the test $\mathbb{1}\{T_{\mathsf{LF}}^{-\mathsf{d}} \ge 0\}$ to the transformed samples with probability mass functions $\tilde{f}, \tilde{g}, \tilde{h}$. Note that taking c_{r} small eonugh shows that

$$\|\tilde{g} - \tilde{h}\|_2^2 - \|\tilde{f} - \tilde{h}\|_2^2 \gtrsim \|\tilde{f} - \tilde{g}\|_2^2$$

for a universal implied constant. Therefore, by Proposition A.1.1 we see that $-\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} \ge c\|\tilde{f} - \tilde{g}\|_2^2 + R$ for some universal constant $c > 0$, where the residual term R can be bounded as

$$\begin{aligned} |R| &= \left|\frac{\|\tilde{f}\|_2^2 - \|\tilde{g}\|_2^2}{n}\right| \\ &\lesssim \frac{\|\tilde{f} - \tilde{g}\|_2}{n\sqrt{k \wedge (m \vee n)}}, \end{aligned}$$

where we used (A.1.13). We have $-\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} \gtrsim \|\tilde{f}-\tilde{g}\|_2^2$ provided $n \gtrsim 1/(\|\tilde{f}-\tilde{g}\|_2\sqrt{k \wedge (m \vee n)}) \asymp \sqrt{\alpha}/\epsilon$, which we assume from here on. Plugging in the bounds derived above, the test $\mathbb{1}\{T_{\mathsf{LF}} \geq 0\}$ on the transformed observations has type-I probability of error bounded by $1/3$ provided

$$\|\tilde{f}-\tilde{g}\|_2^4 \gtrsim \frac{1}{n}\sqrt{\frac{\alpha}{k}}\|\tilde{f}-\tilde{g}\|_2^2 + \frac{1}{m}\frac{\log(k)\alpha}{k}\|\tilde{f}-\tilde{g}\|_2^2 + \frac{\alpha}{k}\left(\frac{1}{nm}+\frac{1}{n^2}\right)$$

for a small enough implied constant on the left. Looking at each term separately yields the sufficient conditions

$$\underbrace{m \gtrsim \frac{\log(k)\alpha}{\epsilon^2}}_{(I)} \qquad \text{and} \qquad n \gtrsim \frac{\sqrt{k\alpha}}{\epsilon^2} \qquad \text{and} \qquad mn \gtrsim \frac{k\alpha}{\epsilon^4}. \tag{A.1.16}$$

The final step is to check that the sufficient conditions in (A.1.16) are implied by what is indicated in the statement of Theorem 2.3.3. Recall from the statement of the Theorem, that it states that

$$m \gtrsim \frac{\log(k)}{\epsilon^2} \qquad \text{and} \qquad n \gtrsim \frac{\sqrt{k\alpha}}{\epsilon^2} \qquad \text{and} \qquad mn \gtrsim \frac{k\log(k)\alpha}{\epsilon^4} \tag{A.1.17}$$

is sufficient to successfully perform the test, where we have replaced the generic $\gtrsim_{\log(k)}$ notation with the precise dependence on $\log(k)$ that we require. Note that the only difference between (A.1.16) and (A.1.17) is the condition on m, that is, the first term in the equations (A.1.16) and (A.1.17). Suppose now that (A.1.17) holds, and let us split this discussion into cases.

1. Suppose $\max\{m,n\} \geq k$. In this case $\alpha = 1$, and (I) is implied by $m \gtrsim \log(k)/\epsilon^2$. For this the first condition of (A.1.17) is clearly sufficient.

2. Suppose $n \leq m \leq k$. In this case $\alpha = k/m$, and (I) is implied by $m \gtrsim \sqrt{k\log(k)}/\epsilon$. By the third condition of (A.1.17) we know that $m^2 n \gtrsim k^2/\epsilon^4$. Using that $n \leq m$, this implies that $m \gtrsim k^{2/3}/\epsilon^{4/3}$, which is clearly sufficient.

3. Suppose $m \leq n \leq k$. In this case $\alpha = k/n$, and (I) is implied by $mn \gtrsim k\log(k)/\epsilon^2$. By the third condition of (A.1.17) we know that $mn^2 \gtrsim k^2\log(k)/\epsilon^4$. After noting that $n \leq k$ we get $mn \gtrsim k\log(k)/\epsilon^4$, which is sufficient.

The diagonal. See the discussion at the end of the proof for $\mathcal{P}_{\mathsf{Db}}$. □

A.2 Lower bounds of Theorem 2.3.2 and 2.3.3

Let $\mathcal{M}(\mathcal{X})$ be the set of all probability measures on some space $\mathcal{X}$, and $\mathcal{P} \subseteq \mathcal{M}(\mathcal{X})$ be some family of distributions. In this section we prove lower bounds for likelihood-free hypothesis

testing problems. For clarity, let us formally state the problem as testing between the hypotheses

$$\begin{aligned} H_0 &= \{\mathbb{P}_{\mathsf{X}}^{\otimes n} \otimes \mathbb{P}_{\mathsf{Y}}^{\otimes n} \otimes \mathbb{P}_{\mathsf{X}}^{\otimes m} : \mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P},\ \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \geq \epsilon\} \\ &\quad \text{versus} \\ H_1 &= \{\mathbb{P}_{\mathsf{X}}^{\otimes n} \otimes \mathbb{P}_{\mathsf{Y}}^{\otimes n} \otimes P_{\mathsf{Y}}^{\otimes m} : \mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P},\ \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \geq \epsilon\}. \end{aligned} \tag{A.2.1}$$

Our strategy for proving lower bounds relies on the following well known result proved in the main text.

Lemma A.2.1. *Take hypotheses $H_0, H_1 \subseteq \mathcal{M}(\mathcal{X})$ and $P_0, P_1 \in \mathcal{M}(\mathcal{X})$ random. Then*

$$\inf_{\psi} \max_{i=0,1} \sup_{P \in H_i} P(\psi \neq i) \geq \frac{1}{2}\left(1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1)\right) - \sum_i \mathbb{P}(P_i \notin H_i),$$

where the infimum is over all tests $\psi : \mathcal{X} \to \{0, 1\}$.

The following will also be used multiple times throughout:

Lemma A.2.2 ([198, Lemmas 2.3 and 2.4])**.** *For any probability measures $\mathbb{P}_0, \mathbb{P}_1$,*

$$\frac{1}{4}\mathsf{H}^4(\mathbb{P}_0, \mathbb{P}_1) \leq \mathsf{TV}^2(\mathbb{P}_0, \mathbb{P}_1) \leq \mathsf{H}^2(\mathbb{P}_0, \mathbb{P}_1) \leq \mathsf{KL}(\mathbb{P}_0 \| \mathbb{P}_1) \leq \chi^2(\mathbb{P}_0 \| \mathbb{P}_1).$$

Note that some of the inequalities in Lemma A.2.2 can be improved, but since such improvements have no effect on our results, we present their simplest available version. The inequalities between TV and H are attributed to Le Cam, while the bound $\mathsf{TV} \leq \sqrt{\mathsf{KL}/2}$ is due to Pinsker. The use of the χ^2-divergence for bounding the total variation distance between mixtures of products was pioneered by Ingster [113], and is sometimes referred to as the *Ingster-trick*.

In our bounds we will also rely on the following simple technical result.

Lemma A.2.3. *Suppose that $a, b, c > 0$ and $N = (N_1, \ldots, N_k) \sim \text{Multinomial}(n, (\frac{1}{k}, \ldots, \frac{1}{k}))$. Then*

$$\mathbb{E}_N \prod_{j \in [k]} (a + b(1 + c)^{N_j}) \leq (a + be^{cn/k})^k.$$

Recall that the necessity of $m \gtrsim n_{\mathsf{HT}}(\epsilon, \mathcal{P})$ and $n \gtrsim n_{\mathsf{GoF}}(\epsilon, \mathcal{P})$ were shown in Proposition 2.3.1. Thus, most of our work lies in obtaining the lower bound on the product mn.

A.2.1 The class $\mathcal{P}_{\mathsf{H}}$

Proposition A.2.4. *For any $\beta > 0, C > 1$ and $d \geq 1$ there exists a finite c independent of ϵ such that*

$$c\{m \geq 1/\epsilon^2, n \geq \epsilon^{-(2\beta + d/2)/\beta}, mn \geq \epsilon^{-2(2\beta + d/2)/\beta}\} \supseteq \mathcal{R}_{\mathsf{LF}}(\epsilon, \mathcal{P}_{\mathsf{H}}(\beta, d, C))$$

for all $\epsilon \in (0, 1)$.

Proof. **Adversarial construction.** Take a smooth function $h : \mathbb{R}^d \to \mathbb{R}$ supported on of $[0,1]^d$ with $\int_{[0,1]^d} h(x)\mathrm{d}x = 0$ and $\int_{[0,1]^d} h(x)^2\mathrm{d}x = 1$. Let $\kappa \geq 1$ be an integer, and for $j \in [\kappa]^d$ define the scaled and translated functions h_j as

$$h_j(x) = \kappa^{d/2} h(\kappa x - j + 1).$$

Then h_j is supported on the cube $[(j-1)/\kappa, j/\kappa]$ and $\int_{[0,1]^d} h_j(x)^2\mathrm{d}x = 1$, where we write $j/\kappa = (j_1/\kappa, \ldots, j_d/\kappa)$. Let $\rho > 0$ be small and for each $\eta \in \{-1, 0, 1\}^{\kappa^d}$ define the function

$$f_\eta(x) = 1 + \rho \sum_{j \in [\kappa]^d} \eta_j h_j(x).$$

In particular, $f_0 = 1$ is the uniform density. Clearly $\int_{[0,1]^d} f_\eta(x)\mathrm{d}x = 1$, and to make it positive we choose ρ, κ such that $\rho\kappa^{d/2}\|h\|_\infty \leq 1/2$. By [10], choosing

$$\rho\kappa^{d/2+\beta} \leq C/(4\|h\|_{\mathcal{C}^{\lfloor\beta\rfloor}} \vee 2\|h\|_{\mathcal{C}^{\lfloor\beta\rfloor+1}}) \tag{A.2.2}$$

ensures that $f_\eta \in \mathcal{P}(\beta, d, C)$. Note also that $\|f_\eta - 1\|_1 = \rho\kappa^{d/2}$. For $\epsilon \in (0,1)$ we set $\kappa \asymp \epsilon^{-1/\beta}$ and $\rho \asymp \epsilon^{(2\beta+d)/(2\beta)}$. These ensure that (A.2.2) and $\mathsf{TV}(f_\eta, f_0) \gtrsim \epsilon$ hold, where as usual the constants may depend on (β, d, C). Noting that $\|\sqrt{f_\eta} - 1\|_2 \asymp \|f_\eta - 1\|_1 \gtrsim \epsilon$, we immediately obtain that $m \gtrsim 1/\epsilon^2$ is necessary for testing, by reduction from binary hypothesis testing (2.3.1). Observe also that for any η, η',

$$\int_{[0,1]^d} f_\eta(x) f_{\eta'}(x)\mathrm{d}x = 1 + \rho^2\langle\eta, \eta'\rangle \tag{A.2.3}$$

which will be used later.

Goodness-of-fit testing. Let η be drawn uniformly at random. We show that $\mathsf{TV}(f_0^{\otimes n}, \mathbb{E} f_\eta^{\otimes n})$ can be made arbitrarily small provided $n \lesssim \epsilon^{-(2\beta+d/2)/\beta}$, which yields a lower bound on n via reduction from goodness-of-fit testing (2.3.3). By Lemma A.2.2 we can focus on bounding the χ^2 divergence. Via Ingster's trick we have

$$\begin{aligned}
\chi^2(\mathbb{E}_\eta[f_\eta^{\otimes n}], f_0^{\otimes n}) + 1 &= \int_{[0,1]^d \times \cdots \times [0,1]^d} \left(\mathbb{E}_\eta \prod_{i=1}^n f_\eta(x_i)\right)^2 \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&= \mathbb{E}_{\eta\eta'} \prod_{i=1}^n \left(\int_{[0,1]^d} f_\eta(x) f_{\eta'}(x)\mathrm{d}x\right),
\end{aligned}$$

where η, η' are i.i.d.. By (A.2.3) and the inequalities $1 + x \leq e^x, \cosh(x) \leq \exp(x^2)$ for all $x \in \mathbb{R}$, we have

$$\begin{aligned}
&= \mathbb{E}_{\eta\eta'} \left(1 + \rho^2\langle\eta, \eta'\rangle\right)^n \\
&\leq \mathbb{E}_{\eta\eta'} \exp(n\rho^2\langle\eta, \eta'\rangle) \\
&= \cosh(n\rho^2)^{\kappa^d} \\
&\leq \exp(n^2\rho^4\kappa^d).
\end{aligned}$$

Thus, goodness-of-fit testing is impossible unless $n \gtrsim \rho^{-2}\kappa^{-d/2} \asymp 1/\epsilon^{(2\beta+d/2)/\beta}$.

Likelihood-free hypothesis testing. We are now ready to show the lower bound on the product mn. Once again $\eta \in \{\pm 1\}^{\kappa^d}$ is drawn uniformly at random and we apply Lemma A.2.1 with the choices $P_0 = f_\eta^{\otimes n} \otimes f_0^{\otimes n} \otimes f_\eta^{\otimes m}$ against $P_1 = f_\eta^{\otimes n} \otimes f_0^{\otimes n+m}$. Let $\mathbb{P}_{0,XYZ}, \mathbb{P}_{1,XYZ}$ denote the joint distribution of the samples X, Y, Z under the measures $\mathbb{E}P_0, \mathbb{E}P_1$ respectively. By Pinsker's inequality and the chain rule we have

$$\begin{aligned}
\mathsf{TV}(\mathbb{P}_{0,XYZ}, \mathbb{P}_{1,XYZ})^2 &= \mathsf{TV}(\mathbb{P}_{0,XZ}, \mathbb{P}_{1,XZ})^2 \\
&\le \mathsf{KL}(\mathbb{P}_{0,XZ} \| \mathbb{P}_{1,XZ}) \\
&= \mathsf{KL}(\mathbb{P}_{0,Z|X} \| \mathbb{P}_{1,Z|X} | \mathbb{P}_{0,X}) + \underbrace{\mathsf{KL}(\mathbb{P}_{0,X} \| \mathbb{P}_{1,X})}_{=0},
\end{aligned}$$

where the last line uses that the marginal of X is equal under both measures. Clearly $\mathbb{P}_{1,Z|X}$ is simply $\mathrm{Unif}([0,1]^d)^{\otimes m}$ and $\mathbb{P}_{0,X}, \mathbb{P}_{0,Z|X}$ have densities $\mathbb{E}_\eta f_\eta^{\otimes n}$ and $\mathbb{E}_{\eta|X} f_\eta^{\otimes m}$ respectively. Given X, let η' be an independent copy of η from the posterior given X. By Ingster's trick we have

$$\begin{aligned}
\mathsf{KL}(\mathbb{P}_{0,Z|X} \| \mathbb{P}_{1,Z|X} | \mathbb{P}_{0,X}) &\le \chi^2(\mathbb{P}_{0,Z|X} \| \mathbb{P}_{1,Z|X} | \mathbb{P}_{0,X}) \\
&= -1 + \mathbb{E}_X \int_{[0,1]^d \times \cdots \times [0,1]^d} \mathbb{E}_{\eta|X} \mathbb{E}_{\eta'|X} \prod_{i=1}^m f_\eta(z_i) f_{\eta'}(z_i) \mathrm{d}z_1 \ldots \mathrm{d}z_m \\
&= -1 + \mathbb{E}_{\eta\eta'}(1 + \rho^2 \langle \eta, \eta' \rangle)^m,
\end{aligned}$$

where the last line uses (A.2.3). Let $N = (N_1, \ldots, N_{\kappa^d})$ be the vector of counts indicating the number of X_i that fall into each bin $\{[(j-1)/\kappa, j/\kappa]\}_{j \in [\kappa]^d}$. Clearly $N \overset{d}{\sim} \mathrm{Multinomial}(n, (\frac{1}{\kappa^d}, \ldots, \frac{1}{\kappa^d}))$. Using that $\eta_j \eta_j'$ depends on only those X_i that fall in bin j and the inequality $1 + x \le \exp(x)$ valid for all $x \in \mathbb{R}$, we can write

$$\begin{aligned}
\chi^2(\mathbb{P}_{0,Z|X} \| \mathbb{P}_{1,Z|X} | \mathbb{P}_{0,X}) + 1 &\le \mathbb{E}_N \mathbb{E}_{\eta\eta'|N} \prod_{j \in [\kappa]^d} \exp(\rho^2 m \eta_j \eta_j') \\
&= \mathbb{E}_N \prod_{j \in [\kappa]^d} \mathbb{E}_{\eta_j \eta_j' | N_j} \exp(\rho^2 m \eta_j \eta_j').
\end{aligned}$$

We now focus on a particular bin j. Define the bin-conditional densities

$$p_\pm = \kappa^d (1 \pm \rho h_j) \mathbb{1}_{[(j-1)/\kappa, j/\kappa]}, \tag{A.2.4}$$

where we drop the dependence on j in the notation. Let $X^{(j)} =: (X_{i_1}, \ldots, X_{i_{N_j}})$ be those X_i that fall in bin j. Note that $\{i_1, \ldots, i_{N_j}\}$ is a uniformly distributed size N_j subset of $[n]$ and given N_j, the density of $X_{i_1}, \ldots, X_{i_{N_j}}$ is $\frac{1}{2}(p_+^{\otimes N_j} + p_-^{\otimes N_j})$. We can calculate

$$\begin{aligned}
\mathbb{P}(\eta_j \eta_j' = 1 | N_j) &= \mathbb{E}_{X^{(j)}|N_j} \mathbb{P}(\eta_j \eta_j' = 1 | X^{(j)}) \\
&= \mathbb{E}_{X^{(j)}|N_j} \left[\mathbb{P}(\eta_j = 1 | X^{(j)})^2 + \mathbb{P}(\eta_j = -1 | X^{(j)})^2 \right] \\
&= \mathbb{E}_{X^{(j)}|N_j} \left[\frac{\frac{1}{4}(p_+^{\otimes N_j})^2 + \frac{1}{4}(p_-^{\otimes N_j})^2}{\frac{1}{4}(p_+^{\otimes N_j} + p_-^{\otimes N_j})^2} \right] \\
&= \frac{1}{2} + \frac{1}{4} \left(\chi^2(p_+^{\otimes N_j} \| \frac{1}{2}(p_+^{\otimes N_j} + p_-^{\otimes N_j})) + \chi^2(p_-^{\otimes N_j} \| \frac{1}{2}(p_+^{\otimes N_j} + p_-^{\otimes N_j})) \right).
\end{aligned}$$

By convexity of the χ^2 divergence in its arguments and tensorization, we have

$$\begin{aligned}\mathbb{P}(\eta_j\eta_j' = 1|N_j) &\leq \frac{1}{2} + \frac{1}{8}\left(\chi^2(p_+^{\otimes N_j}\|p_-^{\otimes N_j}) + \chi^2(p_-^{\otimes N_j}\|p_+^{\otimes N_j})\right) \\ &= \frac{1}{4} + \sum_{\omega\in\{\pm 1\}}\left(\kappa^d\int_{[(j-1)/\kappa, j/\kappa]}\frac{(1+\omega\rho h_j(x))^2}{1-\omega\rho h_j(x)}\mathrm{d}x\right)^{N_j}.\end{aligned}$$

Using that $\rho\|h_j\|_\infty \leq 1/2$ by construction, we have

$$\begin{aligned}\int_{[(j-1)/\kappa, j/\kappa]}\frac{(1+\rho h_j(x))^2}{1-\rho h_j(x)}\mathrm{d}x &= \frac{1}{\kappa^d} + \int_{[(j-1)/\kappa, j/\kappa]}\frac{4\rho^2 h_j^2(x)}{1-\rho h_j(x)}\mathrm{d}x \\ &\leq \frac{1}{\kappa^d} + 8\rho^2.\end{aligned}$$

The same bound is obtained for the other integral term. We get

$$\chi^2(\mathbb{P}_{0,Z|X}\|\mathbb{P}_{1,Z|X}|\mathbb{P}_{0,X})+1 \leq \mathbb{E}_N\prod_{j\in[\kappa]^d}\left(\frac{1}{4}\left(e^{\rho^2 m} - e^{-\rho^2 m}\right)(1 + (1+8\rho^2\kappa^d)^{N_j}) + e^{-\rho^2 m}\right) = (\dagger).$$

The final step is to apply Lemma A.2.3 to pass the expectation through the product. Assuming that $m \vee n \lesssim \rho^{-2} \asymp \epsilon^{-(2\beta+d)/\beta}$ for a small enough implied constant, using the inequalities $e^x \leq 1 + x + x^2, 1 - x \leq e^{-x} \leq 1 - x + x^2/2$ valid for all $x \in [0,1]$, and Lemma A.2.3, we obtain

$$\begin{aligned}(\dagger) &\leq (e^{-\rho^2 m} + \frac{1}{4}\left(e^{\rho^2 m} - e^{-\rho^2 m}\right)(1 + e^{8\rho^2 n}))^{\kappa^d} \\ &\leq (1 + c\rho^4 mn)^{\kappa^d} \\ &\leq \exp(c\rho^4\kappa^d mn)\end{aligned}$$

for a universal constant $c > 0$. Therefore, if $m \vee n \lesssim \epsilon^{-(2\beta+d)/\beta}$ likelihood-free hypothesis testing is impossible unless $mn \gtrsim \rho^{-4}\kappa^{-d} \asymp 1/\epsilon^{2(2\beta+d/2)/\beta}$.

Suppose now that $m \vee n \gtrsim \epsilon^{-(2\beta+d)/\beta}$ instead. We have two cases:

1. If $n \gtrsim \epsilon^{-(2\beta+d)/\beta}$ then from Proposition A.1.3 we know that $m \asymp 1/\epsilon^2$ is enough for achievability. However, by the first part of the proof we know that $m \gtrsim 1/\epsilon^2$ must always hold, which provides the matching lower bound in this case.

2. If $m \gtrsim \epsilon^{-(2\beta+d)/\beta}$ then we can assume $m \gtrsim n$ also holds, otherwise the first case above would apply. From the goodness-of-fit testing lower bound we know that $n \gtrsim \epsilon^{-(2\beta+d/2)/\beta}$ must always hold, and from Proposition A.1.3 we know that $(m,n) \asymp (\epsilon^{-(2\beta+d/2)/\beta}, \epsilon^{-(2\beta+d/2)/\beta})$ is achievable, so we get matching bounds in this case too.

Summarizing, we've shown that for succesful testing $m \gtrsim 1/\epsilon^2, n \gtrsim 1/\epsilon^{(2\beta+d/2)/\beta}$ and $mn \gtrsim \epsilon^{-2(2\beta+d/2)/\beta}$ must hold, which concludes our proof. □

A.2.2 The class $\mathcal{P}_{\mathsf{G}}$

Proposition A.2.5. *For any $s, C > 0$ there exists a finite constant c independent of ϵ such that*

$$c\{m \geq 1/\epsilon^2, n \geq \epsilon^{-(2s+1/2)/s}, mn \geq \epsilon^{-2(2s+1/2)/s}\} \supseteq \mathcal{R}_{\mathsf{LF}}(\epsilon, \mathcal{P}_{\mathsf{G}}(s, C))$$

for all $\epsilon \in (0, 1)$.

Proof. **Adversarial construction.** Let $\gamma \in \ell^1$ be a non-negative sequence, and let $\theta \sim \otimes_{k=1}^{\infty} \mathcal{N}(0, \gamma_k)$. Define the random measure $\mu_\theta = \otimes_{j=1}^{\infty} \mathcal{N}(\theta_j, 1)$. Let $\epsilon \in (0, 1)$ be given. For our proofs we use

$$\gamma_k = \begin{cases} c_1 \epsilon^{(2s+1)/s} & \text{for } 1 \leq k \leq c_2 \epsilon^{-1/s} \\ 0 & \text{otherwise} \end{cases} \tag{A.2.5}$$

for appropriate constants c_1, c_2. Recall our definition of the Sobolev ellipsoid $\mathcal{E}(s, C)$ with associated sobolev norm $\|\cdot\|_s$. We have

$$(\mathbb{E}\|\theta\|_s)^2 \leq \mathbb{E} \sum_{j=1}^{\infty} j^{2s} \theta_i^2 = \|\sqrt{\gamma}\|_s^2 = c_1 \epsilon^{(2s+1)/s} \sum_{j=1}^{c_2 \epsilon^{-1/s}} j^{2s} \leq c_1 c_2^{2s+1}$$

$$\mathsf{TV}(\mathbb{P}_\gamma, \mathbb{P}_0) \geq \frac{1 \wedge \|\theta\|_2}{200},$$

where last line holds by [61, Theorem 1.2].

First, we need to verify that our construction is valid, that is, that $\mathbb{P}_\gamma \in \mathcal{P}_{\mathsf{G}}(s, C)$ and $\mathsf{TV}(\mathbb{P}_\gamma, \mathbb{P}_0) \geq \epsilon$ with high probability. For standard Gaussian $Z \sim \mathcal{N}(0, 1)$ it holds that

$$\mathbb{E} \exp(\lambda(Z^2 - 1)) \leq \exp(2\lambda^2)$$

for all $|\lambda| \leq 1/4$. Therefore, for a sequence of independent standard Gaussians $Z_1, Z_2, \ldots$ we get

$$\mathbb{E} \exp(\lambda \sum_{j=1}^{\infty} \gamma_j (Z_j^2 - 1)) \leq \exp(2\lambda^2 \|\gamma\|_2^2)$$

for all $|\lambda| \leq \min_j (4\gamma_j)^{-1} = c_1^{-1} \epsilon^{-(2s+1)/s}/4$. Assuming that $c_1 \epsilon^{(2s+1)/s} \leq \|\gamma\|_2$, standard sub-Exponential concentration bounds imply that there exists a universal constant $c_3 > 0$ such that

$$\mathbb{P}(\|\theta\|_2^2 - \mathbb{E}\|\theta\|_2^2 \leq -t) \leq \exp(-\frac{c_3 t}{\|\gamma\|_2})$$

for all $t \geq 0$. Since $\mathbb{E}\|\theta\|_2^2 = \|\gamma\|_1 = c_1 c_2 \epsilon^2$, and $\|\gamma\|_2^2 = c_2 c_1^2 \epsilon^{\frac{4s+1}{s}}$, we can set $t = \frac{1}{2}\|\theta\|_2^2$ to get

$$\mathbb{P}(\|\theta\|_2^2 \leq \frac{1}{2} c_1 c_2 \epsilon^2) \leq \exp(-\frac{1}{2} c_3 \sqrt{c_2} \epsilon^{-1/(2s)}).$$

Now choose c_1 and c_2 to satisfy

$$100 c_1 c_2^{2s+1} = C \qquad \text{and} \qquad c_1 c_2 = 2. \tag{A.2.6}$$

and ϵ small enough to satisfy

$$c_1\epsilon^{(2s+1)/s} \le \|\gamma\|_2 = \sqrt{c}_1 c_1 \epsilon^{(2s+1/2)/s} \qquad \text{and} \qquad \frac{1}{2}c_3\sqrt{c}_2\epsilon^{-1/(2s)} \ge \log(100).$$

Long story short, these conditions ensure that $\mathbb{P}(\mu_\gamma \in \mathcal{P}_{\mathsf{G}}(s, C), \mathsf{TV}(\mu_\gamma, \mu_0) \ge \epsilon) \ge 0.98$ for all ϵ small enough in terms of C and s, and therefore we can proceed to computation using Lemma A.2.1.

Note that we immediately get the binary hypothesis testing lower bound $m \gtrsim 1/\epsilon^2$ via our reduction (2.3.3), as $\mathsf{H}(\mu_0, \mu_{\sqrt{\gamma}}) \asymp \mathsf{TV}(\mu_0, \mu_{\sqrt{\gamma}}) = \sqrt{2}\epsilon$ by Lemma 2.3.5 and the choice (A.2.6).

Goodness-of-fit testing. We show that $\mathsf{TV}(\mu_0^{\otimes n}, \mathbb{E}\mu_\gamma^{\otimes n})$ can be made arbitrarily small as long as $n \lesssim 1/\epsilon^{(2s+1/2)/s}$, which yields a lower bound on n via reduction from goodness-of-fit testing (2.3.3). Let us compute the distribution $\mathbb{E}\mu_\gamma^{\otimes n}$. By independence clearly $\mathbb{E}\mu_\gamma^{\otimes n} = \otimes_{k=1}^\infty \mathbb{E}_{\theta\sim\mathcal{N}(0,\gamma_k)}\mathcal{N}(\theta, 1)^{\otimes n}$. Focusing on the inner term and and dropping the subscript k, for the density we have

$$\mathbb{E}_{\theta\sim\mathcal{N}(0,\gamma)}\left[\frac{1}{(2\pi)^{n/2}}\exp(-\frac{1}{2}\sum_{j=1}^n(x_j-\theta)^2)\right] \propto \exp(-\frac{\|x\|_2^2}{2})\mathbb{E}\exp(-\frac{n}{2}(\theta^2 - 2\theta\bar{x})),$$

where we write $\bar{x} =: \frac{1}{n}\sum_j x_j$. Looking at just the term involving θ, we have

$$\mathbb{E}\exp(-\frac{n}{2}(\theta^2-2\theta\bar{x})) \propto \int \exp(-\frac{1}{2}(\theta^2(n+\frac{1}{\gamma}) - 2\theta n\bar{x}))\mathrm{d}\theta \propto \exp(\frac{1}{2}\frac{n^2\bar{x}^2}{n+\frac{1}{\gamma}}).$$

Putting everything together, we see that $\mathbb{E}\mu_\gamma^{\otimes n} = \otimes_{k=1}^\infty \mathcal{N}(0, (\mathsf{Id}_n - \frac{\gamma_k}{1+n\gamma_k}\mathbb{1}_n\mathbb{1}_n^\mathsf{T})^{-1})$. Thus, using Lemma A.2.2 we obtain

$$\begin{aligned}
\mathsf{TV}^2(\mu_0^{\otimes n}, \mathbb{E}\mu_\gamma^{\otimes n}) &\le \sum_{k=1}^\infty \mathsf{KL}(\mathcal{N}(0, \mathsf{Id}_n)\|\mathcal{N}(0, (\mathsf{Id}_n - \frac{\gamma_k}{1+n\gamma_k}\mathbb{1}_n\mathbb{1}_n^\mathsf{T})^{-1})) \\
&= \frac{1}{2}\sum_{k=1}^\infty\left(-\frac{n\gamma_k}{n\gamma_k+1} + \log(1+n\gamma_k)\right) \\
&\le \frac{1}{2}\sum_{k=1}^\infty \frac{n^2\gamma_k^2}{1+n\gamma_k} \lesssim \sum_{k=1}^\infty n^2\gamma_k^2.
\end{aligned}$$

Taking γ as in (A.2.5) gives

$$\mathsf{TV}^2(\mu_0^{\otimes n}, \mathbb{E}\mu_\gamma^{\otimes n}) \lesssim n^2\epsilon^{2(2s+1/2)/s}.$$

Thus, goodness-of-fit testing is impossible unless $n \gtrsim 1/\epsilon^{(2s+1/2)/s}$ as desired.

Likelihood-free hypothesis testing. We apply Lemma A.2.1 with measures $P_0 = \mu_\gamma^{\otimes n} \otimes \mu_0^{\otimes n} \otimes \mu_\gamma^{\otimes m}$ and $P_1 = \mu_\gamma^{\otimes n} \otimes \mu_0^{\otimes n} \otimes \mu_0^{\otimes m}$. By an analogous calculation to that in the

previous part, we obtain

$$\mathbb{E}P_0 = \otimes_{k=1}^{\infty}\mathcal{N}\left(0, \left(\mathsf{Id}_{2n+m} - \frac{1}{n+m+\frac{1}{\gamma_k}}\begin{pmatrix}\mathbb{1}_n\mathbb{1}_n^{\mathsf{T}} & 0 & \mathbb{1}_n\mathbb{1}_m^{\mathsf{T}} \\ 0 & 0 & 0 \\ \mathbb{1}_m\mathbb{1}_n^{\mathsf{T}} & 0 & \mathbb{1}_m\mathbb{1}_m^{\mathsf{T}}\end{pmatrix}\right)^{-1}\right) =: \otimes_{k=1}^{\infty}\mathcal{N}(0, \Sigma_{0k})$$

$$\mathbb{E}P_1 = \otimes_{k=1}^{\infty}\mathcal{N}\left(0, \left(\mathsf{Id}_{2n+m} - \frac{1}{n+\frac{1}{\gamma_k}}\begin{pmatrix}\mathbb{1}_n\mathbb{1}_n^{\mathsf{T}} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0\end{pmatrix}\right)^{-1}\right) =: \otimes_{k=1}^{\infty}\mathcal{N}(0, \Sigma_{1k}).$$

By the Sherman-Morrison formula, we have

$$\Sigma_{0k} = \mathsf{Id}_{2n+m} + \gamma_k\begin{pmatrix}\mathbb{1}_n\mathbb{1}_n^{\mathsf{T}} & 0 & \mathbb{1}_n\mathbb{1}_m^{\mathsf{T}} \\ 0 & 0 & 0 \\ \mathbb{1}_m\mathbb{1}_n^{\mathsf{T}} & 0 & \mathbb{1}_m\mathbb{1}_m^{\mathsf{T}}\end{pmatrix}$$

Therefore, by Pinsker's inequality and the closed form expression for the KL-divergence between centered Gaussians, we obtain

$$\begin{aligned}\mathsf{TV}^2(\mathbb{E}P_0, \mathbb{E}P_1) &\le \mathsf{KL}(\mathbb{E}P_0 \| \mathbb{E}P_1) \\ &= \frac{1}{2}\sum_{k=1}^{\infty}\left(\gamma_k m - \log\left(1 + \frac{\gamma_k m}{\gamma_k(n+m)+1}\right)\right).\end{aligned}$$

Once again we choose γ as in (A.2.5). Using the inequality $\log(1+x) \ge x - x^2$ valid for all $x \ge 0$ we obtain

$$\mathsf{TV}^2(\mathbb{E}P_0, \mathbb{E}P_1) \lesssim \epsilon^{-2(2s+1/2)/s}(m^2 + mn).$$

Therefore, likelihood-free hypothesis testing is impossible unless $m \gtrsim \epsilon^{-(2s+1/2)/s}$ or $nm \gtrsim \epsilon^{-2(2s+1/2)/s}$. Note that we already have the lower bound $n \gtrsim \epsilon^{-(2s+1/2)/s}$ by reduction from goodness-of-fit testing (2.3.3), so that $m \gtrsim \epsilon^{-(2s+1/2)/s}$ automatically implies $nm \gtrsim \epsilon^{-2(2s+1/2)/s}$. Combining everything we get the desired bounds. □

A.2.3 The classes $\mathcal{P}_{\mathsf{Db}}$ and $\mathcal{P}_{\mathsf{D}}$

Our first result in this section derives tight minimax lower bounds for the class $\mathcal{P}_{\mathsf{Db}}$. Since $\mathcal{P}_{\mathsf{D}} \supset \mathcal{P}_{\mathsf{Db}}$ these lower bounds immediately carry over to the larger class. However, to get tight lower bounds for all regimes for $\mathcal{P}_{\mathsf{D}}$, we have to prove additional results in Propositions A.2.7 and A.2.9 below.

Proposition A.2.6. *For any $C > 1$ there exists a finite constant c independent of ϵ and k, such that*

$$c\{m \ge 1/\epsilon^2, n \ge \sqrt{k}/\epsilon^2, mn \ge k/\epsilon^4\} \supseteq \mathcal{R}_{\mathsf{LF}}(\epsilon, \mathcal{P}_{\mathsf{Db}}(k, C)) \supseteq \mathcal{R}_{\mathsf{LF}}(\epsilon, \mathcal{P}_{\mathsf{D}}(k))$$

for all $\epsilon \in (0,1)$ and $k \ge 2$.

Proof. The second inclusion is trivial. For the first inclusion we proceed analogously to the case of $\mathcal{P}_{\mathsf{H}}$.

Adversarial construction. Let k be an integer and $\epsilon \in (0,1)$. For $\eta \in \{-1,1\}^k$ define the distribution p_η on $[2k]$ by

$$p_\eta(2j-1) = \frac{1}{2k}(1+\eta_j\epsilon)$$
$$p_\eta(2j) = \frac{1}{2k}(1-\eta_j\epsilon),$$

for $j \in [k]$. Clearly $\mathsf{H}(p_\eta, p_0) \asymp \mathsf{TV}(p_\eta, p_0) = \epsilon$, where $p_0 = \mathrm{Unif}[2k]$, so that by reduction from binary hypothesis testing (2.3.3) we get the lower bound $m \gtrsim 1/\epsilon^2$. Observe also that for any $\eta, \eta' \in \{\pm 1\}^k$,

$$\sum_{j\in[2k]} p_\eta(j)p_{\eta'}(j) = \frac{1}{2k}(1+\frac{\epsilon^2\langle\eta,\eta'\rangle}{k}). \tag{A.2.7}$$

Goodness-of-fit testing. Let η be uniformly random. We show that $\mathsf{TV}(p_0^{\otimes n}, \mathbb{E}p_\eta^{\otimes n})$ can be made arbitrarily small as long as $n \lesssim \sqrt{k}/\epsilon^2$, which yields the corresponding lower bound on n by reduction from goodness-of-fit testing (2.3.3). Once again, by Lemma A.2.2 we focus on the χ^2 divergence. We have

$$\begin{aligned}\chi^2(\mathbb{E}p_\eta^{\otimes n}\|p_0^{\otimes n}) + 1 &= (2k)^n \sum_{j\in[2k]^n} \mathbb{E}_{\eta\eta'} \prod_{i=1}^n p_\eta(j_i)p_{\eta'}(j_i) \\ &= \mathbb{E}_{\eta\eta'}(1+\frac{\epsilon^2\langle\eta,\eta'\rangle}{k})^n \\ &\le \exp(n^2\epsilon^4/k)\end{aligned}$$

where the penultimate line follows from (A.2.7) and the last line via the same argument as in A.2.1. Thus, goodness-of-fit testing is impossible unless $n \gtrsim \sqrt{k}/\epsilon^2$.

Likelihood-free hypothesis testing. We apply Lemma A.2.1 with the two random measures $P_0 = p_\eta^{\otimes n} \otimes p_0^{\otimes n} \otimes p_\eta^{\otimes m}$ and $P_1 = p_\eta^{\otimes n} \otimes p_0^{\otimes(n+m)}$. Analogously to the case of $\mathcal{P}_{\mathsf{H}}$, let $\mathbb{P}_{0,XYZ}, \mathbb{P}_{1,XYZ}$ respectively denote the distribution of the observations X, Y, Z under $\mathbb{E}P_0, \mathbb{E}P_1$ respectively. As for $\mathcal{P}_{\mathsf{H}}$, we have

$$\begin{aligned}\mathsf{TV}^2(\mathbb{P}_{0,XYZ}, \mathbb{P}_{1,XYZ}) &\le \mathsf{KL}(\mathbb{P}_{0,XYZ}\|\mathbb{P}_{1,XYZ}) \\ &\le \mathsf{KL}(\mathbb{P}_{0,Z|X}\|\mathbb{P}_{1,Z|X}|\mathbb{P}_{0,X}).\end{aligned}$$

For any X the distribution $\mathbb{P}_{1,Z|X}$ is uniform, and $\mathbb{P}_{0,Z|X}, \mathbb{P}_{0,X}$ have pmf $\mathbb{E}_{\eta|X}p_\eta^{\otimes m}$ and $\mathbb{E}_\eta p_\eta^{\otimes n}$ respectively. Once again, by Lemma A.2.2 we may turn our attention to the χ^2-divergence. Given X, let η' have the same distribution as η and be independent of it. Then

$$\begin{aligned}\chi^2(\mathbb{P}_{0,Z|X}\|\mathbb{P}_{1,Z|X}|\mathbb{P}_{0,X}) + 1 &= (2k)^m\mathbb{E}_X \sum_{j\in[2k]^m} \mathbb{E}_{\eta|X}\mathbb{E}_{\eta'|X} \prod_{i=1}^n p_\eta(j_i)p_{\eta'}(j_i) \\ &= \mathbb{E}_{\eta\eta'}(1+\frac{\epsilon^2\langle\eta,\eta'\rangle}{k})^m \\ &\le \mathbb{E}_{\eta\eta'} \prod_{j\in[k]} \exp(\frac{\epsilon^2 m\eta_j\eta'_j}{k}),\end{aligned}$$

where we used Lemma A.2.7. Let $N = (N_1, \dots, N_k)$ be the vector of counts indicating the number of the $X_1, \dots, X_n$ that fall into the bins $\{2j-1, 2j\}$ for $j \in [k]$. Clearly $N \sim \mathrm{Mult}(n, (\frac{1}{k}, \dots, \frac{1}{k}))$. Let us focus on a specific bin $\{2j-1, 2j\}$ and define the bin-conditional pmf

$$p_\pm(x) = \begin{cases} \frac{1}{2}(1 \pm \epsilon) & \text{if } x = 2j-1, \\ \frac{1}{2}(1 \mp \epsilon) & \text{if } x = 2j \\ 0 & \text{otherwise,} \end{cases}$$

where we drop the dependence on j in the notation. Let $X_{i_1}, \dots, X_{i_{N_j}}$ be the N_j observations falling in $\{2j-1, 2j\}$. Given N_j, the pmf of $X_{i_1}, \dots, X_{i_{N_j}}$ is $\frac{1}{2}(p_+^{\otimes N_j} + p_-^{\otimes N_j})$. We have $\eta_j \eta_j' \in \{\pm 1\}$ almost surely, and analogously to Section A.2.1 we may compute

$$\begin{aligned} \mathbb{P}(\eta_j \eta_j' = 1 | N_j) &= \mathbb{E}_{X|N_j} \mathbb{P}(\eta_j \eta_j' = 1 | X) \\ &= \mathbb{E}_{X|N_j} \left[\mathbb{P}(\eta_j = 1|X)^2 + \mathbb{P}(\eta_j = -1|X)^2 \right] \\ &= \frac{1}{2} + \frac{1}{4} \left(\chi^2(p_+^{\otimes N_j} \| \frac{1}{2}(p_+^{\otimes N_j} + p_-^{\otimes N_j})) + \chi^2(p_-^{\otimes N_j} \| \frac{1}{2}(p_+^{\otimes N_j} + p_-^{\otimes N_j}) \right) \\ &\le \frac{1}{2} + \frac{1}{8} \left(\chi^2(p_-^{\otimes N_j} \| p_+^{\otimes N_j}) + \chi^2(p_+^{\otimes N_j} \| p_-^{\otimes N_j}) \right). \end{aligned}$$

We can bound the two χ^2-divergences by

$$\begin{aligned} \chi^2(p_\pm^{\otimes N_j} \| p_\mp^{\otimes N_j}) + 1 &= \left(\frac{1 + \frac{3}{2}\epsilon^2}{1 - \epsilon^2} \right)^{N_j} \\ &\le (1 + 3\epsilon^2)^{N_j}, \end{aligned}$$

provided $\epsilon \le c$ for some universal constant $c > 0$. Using Lemma A.2.3, we obtain the bound

$$\begin{aligned} &\mathbb{E}_N \prod_{j \in [k]} \mathbb{E}_{\eta\eta'|N_j} \exp(\frac{\epsilon^2 m \eta_j \eta_j'}{k}) \\ &\qquad \le \mathbb{E}_N \prod_{j \in [k]} \left(\frac{1}{2}(\exp(\frac{\epsilon^2 m}{k}) - \exp(-\frac{\epsilon^2 m}{k}))(1 + (1+2\epsilon^2)^{N_j}) + \exp(-\frac{\epsilon^2 m}{k}) \right) \\ &\qquad \le \left(\frac{1}{2}(\exp(\frac{\epsilon^2 m}{k}) - \exp(-\frac{\epsilon^2 m}{k}))(1 + \exp(\frac{2\epsilon^2 n}{k})) + \exp(-\frac{\epsilon^2 m}{k}) \right)^k. \end{aligned}$$

Now, under the assumption that $m \vee n \lesssim k/\epsilon^2$ for some small enough implied constant, the above can be further bounded by

$$\begin{aligned} &\le (1 + c\frac{\epsilon^4 mn}{k^2})^k \\ &\le \exp(\frac{c\epsilon^4 mn}{k}), \end{aligned}$$

for a universal constant $c > 0$. In other words, for $n \vee m \lesssim k/\epsilon^2$ likelihood-free hypothesis testing is impossible unless $mn \gtrsim k/\epsilon^4$. The treatment of the case $m \vee n \gtrsim k/\epsilon^2$ is straightforward, and entirely analogous to our discussion at the end of the proof of Proposition A.2.4, so we won't repeat it here. This completes the proof. □

This takes care of the class $\mathcal{P}_{\mathsf{Db}}$. To prove tight bounds for $\mathcal{P}_{\mathsf{D}}$ in the large k regime, we have to work harder. Our second lower bound, Proposition A.2.7 below, proves tight bounds in the regime $n \leq m$ and follows by reduction to two-sample testing Proposition 2.3.1.

Proposition A.2.7. *There exists a finite constant c independent of ϵ and k,*

$$c\{m \geq 1/\epsilon^2,\, n^2 m \geq k^2/\epsilon^4,\, n \leq m\} \supseteq \mathcal{R}_{\mathsf{LF}}(\epsilon, \mathsf{TV}, \mathcal{P}_{\mathsf{D}}) \cap \mathbb{N}^2_{n \leq m}$$

for all $k \geq 2, \epsilon \in (0,2)$, where $\mathbb{N}^2_{n\leq m} = \{(n,m) \in \mathbb{N}^2 : n \leq m\}$.

Proof. Follows from (2.3.6) and the lower bound construction in [25]. □

Valiant's wishful thinking theorem.

For our third and final lower bound, which is tight in the regime $m \leq n$, we apply a method developed by Valiant, which we describe below.

Definition 9. *For distributions $p_1, \ldots, p_\ell$ on $[k]$ and $(n_1, \ldots, n_\ell) \in \mathbb{N}^\ell$, we define the $(n_1, \ldots, n_\ell)$-based moments of $(p_1, \ldots, p_\ell)$ as*

$$m(a_1, \ldots, a_\ell) = \sum_{i=1}^{k} \prod_{j=1}^{\ell} (n_j p_j(i))^{a_j}$$

for $(a_1, \ldots, a_\ell) \in \mathbb{N}^\ell$.

Let $p^+ = (p_1^+, \ldots, p_\ell^+)$ and $p^- = (p_1^-, \ldots, p_\ell^-)$ be ℓ-tuples of distributions on $[k]$ and suppose we observe samples $\{X^{(i)}\}_{i\in[\ell]}$, where the number of observations in $X^{(i)}$ is $\mathrm{Poi}(n_i)$. Let $H^\pm$ denote the hypothesis that the samples came from $p^\pm$, up to an arbitrary relabeling of the alphabet $[k]$. It can be shown that to test H^+ against H^-, we may assume without loss of generality that our test is invariant under relabeling of the support, or in other words, is a function of the *fingerprints*. The fingerprint f of a sample $\{X^{(i)}\}_{i\in[\ell]}$ is the function $f : \mathbb{N}^\ell \to \mathbb{N}$ which given $(a_1, \ldots, a_\ell) \in \mathbb{N}^\ell$ counts the number of bins in $[k]$ which have exactly a_i occurences in the sample $X^{(i)}$.

Theorem A.2.8 ([200, Wishful thinking theorem])**.** *Suppose that $|p_i^\pm|_\infty \leq \eta/n_i$ for all $i \in [\ell]$ for some $\eta > 0$, and let m^+ and m^- denote the $(n_1, \ldots, n_\ell)$-based moments of p^+, p^- respectively. Let $f^\pm$ denote the distribution of the fingerprint under $H^\pm$ respectively. Then*

$$\mathsf{TV}(f^+, f^-) \leq 2(e^{\eta\ell} - 1) + e^{\ell(\eta/2 + \log 3)} \sum_{a \in \mathbb{N}^\ell} \frac{|m^+(a) - m^-(a)|}{\sqrt{1 + m^+(a) \vee m^-(a)}}.$$

Proof. The proof is a straightforward adaptation of [200] and thus we omit it. □

Although Theorem A.2.8 assumes a random (Poisson distributed) number of samples, the results carry over to the deterministic case with no modification, due to the sub-exponential concentration of the Poisson distribution. We are ready to prove our likelihood-free hypothesis testing lower bound using Theorem A.2.8.

Proposition A.2.9. *There exists a finite constant c independent of ϵ and k, such that*

$$c\{m \geq 1/\epsilon^2, n^2 m \geq k^2/\epsilon^4,\, m \leq n\} \supseteq \mathcal{R}_{\mathsf{LF}}(\epsilon, \mathsf{TV}, \mathcal{P}_{\mathsf{D}}) \cap \mathbb{N}^2_{m \leq n}$$

for all $\epsilon \in (0,1)$ and $k \geq 2$, where $\mathbb{N}^2_{m\leq n} = \{(n,m) \in \mathbb{N}^2 : m \leq n\}$.

Proof. We focus on the regime $n \leq k$, as otherwise the result is subsumed by Proposition A.2.6. Suppose that $\epsilon \in (0, 1/2)$, $\eta = 0.01$ (say) and $n/\eta \leq k/2$. Define $\gamma = n/\eta$ and let p, q be pmfs on $[k]$ with weight $(1-\epsilon)/\gamma$ on $[\gamma]$ and $k/4$ light elements with weight $4\epsilon/k$ on $[k/2, 3k/4]$ and $[3k/4, k]$ respectively. To apply Valiant's wishful thinking theorem, we take $p^+ = (p, q, p)$ and $p^- = (p, q, q)$ with corresponding hypotheses $H^{\pm}$. The (n, n, m)-based moments of $p^{\pm}$ are given by

$$\frac{1}{n^{a+b}m^c} m^+(a,b,c) = \begin{cases} k & \text{if } a+c = 0, b = 0 \\ \left(\frac{1-\epsilon}{\alpha}\right)^{a+b+c} \alpha + \left(\frac{4\epsilon}{k}\right)^{a+b+c} \frac{k}{4} & \text{if } a+c = 0 \text{ xor } b = 0 \\ \left(\frac{1-\epsilon}{\alpha}\right)^{a+b+c} \alpha & \text{if } a+c \geq 1, b \geq 1, \end{cases}$$

$$\frac{1}{n^{a+b}m^c} m^-(a,b,c) = \begin{cases} k & \text{if } a = 0, b+c = 0 \\ \left(\frac{1-\epsilon}{\alpha}\right)^{a+b+c} \alpha + \left(\frac{4\epsilon}{k}\right)^{a+b+c} \frac{k}{4} & \text{if } a = 0 \text{ xor } b+c = 0 \\ \left(\frac{1-\epsilon}{\alpha}\right)^{a+b+c} \alpha & \text{if } a \geq 1, b+c \geq 1. \end{cases}$$

By the wishful thinking theorem we know that

$$\mathsf{TV}(f^+, f^-) \leq 0.061 + 27.41 \sum_{a,b,c \in \mathbb{N}} \frac{|m^+(a,b,c) - m^-(a,b,c)|}{\sqrt{1 + \max(m^+, m^-)}}.$$

Let us consider the possible values of $|m^+(a,b,c) - m^-(a,b,c)|$. It is certainly zero if $a \wedge b \geq 1$ or $a = b = c = 0$. Suppose that $a = 0$ so that necessarily $b + c \geq 1$. Then

$$\frac{1}{n^b m^c} |m^+(0,b,c) - m^-(0,b,c)| = \left(\frac{4\epsilon}{k}\right)^{b+c} \frac{k}{4} \mathbb{1}(b \wedge c \geq 1).$$

Using the symmetry between a and b and that $1 + m^+ \vee m^- \geq n^b m^c ((1-\epsilon)/\gamma)^{b+c} \gamma$ (for $m^+ \neq m^-$), we can bound the infinite sum above as

$$\begin{aligned} &\lesssim \sum_{b,c \geq 1} \frac{n^b m^c k^{1-(b+c)} \epsilon^{b+c}}{\sqrt{n^b m^c \gamma^{1-(b+c)} (1-\epsilon)^{b+c}}} \\ &\lesssim \sum_{b,c \geq 1} n^{b/2} m^{c/2} \left(\frac{\sqrt{\gamma}}{k}\right)^{b+c-1} \epsilon^{b+c} \end{aligned}$$

Plugging in $\gamma = n/\eta \asymp n$, and using $m \le n \le k$, we obtain

$$\begin{aligned}
\mathsf{TV}(f^+, f^-) - 0.061 &\lesssim \sum_{b,c\ge 1} n^{b+\frac{c}{2}-\frac{1}{2}} m^{c/2} \frac{1}{k^{b+c-1}} \epsilon^{b+c} \\
&= \frac{n\sqrt{m}\epsilon^2}{k} \sum_{b,c\ge 0} \left(\frac{n}{k}\right)^{b+\frac{c}{2}} \left(\frac{m}{k}\right)^{\frac{c}{2}} \epsilon^{b+c} \\
&\le \frac{n\sqrt{m}\epsilon^2}{k} \sum_{b,c\ge 0} \epsilon^{b+c} \\
&\lesssim \frac{n\sqrt{m}\epsilon^2}{k},
\end{aligned}$$

where we use that $\epsilon < 1/2$. Thus, likelihood-free hypothesis testing is impossible for $m \le n$ unless $n^2 m \gtrsim k^2/\epsilon^4$. □

A.3 Proof of Theorem 2.3.6

A.3.1 Upper bound

We deduce the upper bound by applying the corresponding result for $\mathcal{P}_{\mathsf{D}}$ as a black-box procedure.

Theorem A.3.1 ([54]). *For a constant independent of ϵ and k,*

$$n_{\mathsf{GoF}}(\epsilon, \mathsf{H}, \mathcal{P}_{\mathsf{D}}) \asymp \sqrt{k}/\epsilon^2.$$

Write $\mathcal{G}_\ell$ for the regular grid of size ℓ^d on $[0,1]^d$ and let P_ℓ denote the L^2-projector onto the space of functions piecewise constant on the cells of $\mathcal{G}_\ell$. For convenience let us re-state Proposition 2.4.3.

Proposition A.3.2. *For any $\beta \in (0,1]$, $C > 1$ and $d \ge 1$ there exists a constant $c > 0$ such that*

$$c\mathsf{H}(f,g) \le \mathsf{H}(P_\kappa f, P_\kappa g) \le \mathsf{H}(f,g)$$

holds for any $f, g \in \mathcal{P}_{\mathsf{H}}(\beta, d, C)$, provided we set $\kappa = (c\epsilon)^{-2/\beta}$.

With the above approximation result, the proof of Theorem 2.3.6 is straightforward.

Proof of Theorem 2.3.6. Suppose we are testing goodness-of-fit to $f_0 \in \mathcal{P}_{\mathsf{H}}$ based on an i.i.d. sample $X_1, \dots, X_n$ from $f \in \mathcal{P}_{\mathsf{H}}$. Take $\kappa \asymp \epsilon^{-2/\beta}$ and bin the observations on $\mathcal{G}_\kappa$, denoting the pmf of the resulting distribution as p_f. Then, under the alternative hypothesis that $\mathsf{H}(f, f_0) \ge \epsilon$, by Proposition 2.4.3

$$\epsilon \lesssim \mathsf{H}(P_\kappa f_0, P_\kappa f) = \mathsf{H}(p_{f_0}, p_f).$$

In particular, applying the algorithm achieving the upper bound in Theorem A.3.1 to the binned observations, we see that $n \gtrsim \sqrt{\kappa^d}/\epsilon^2 = \epsilon^{-(2\beta+d)/\beta}$ samples suffice. □

A.3.2 Lower bound

The proof is extremely similar to the TV case, except we put the perturbations at density level ϵ^2 instead of 1.

Proof. Let $\phi : [0,1] \to [0,1]$ be a smooth function such that $\phi(x) = 0$ for $x \le 1/3$ and $\phi(x) = 1$ for $x \ge 2/3$. Let $h : \mathbb{R}^d \to \mathbb{R}$ be smooth, supported in $[0,1]^d$, and satisfy $\int_{[0,1]^d} h(x)\mathrm{d}x = 0$ and $\int_{[0,1]^d} h(x)^2\mathrm{d}x = 1$. Given $\epsilon \in (0,1)$ let

$$f_0(x) = \epsilon^2 + \frac{\phi(x_1)}{\|\phi\|_1}(1-\epsilon^2),$$

which is a density on $[0,1]^d$. For a large integer κ and $j \in [\kappa/3] \times [\kappa]^{d-1}$ let

$$h_j(x) = \kappa^{d/2} h(\kappa x - j + 1)$$

for $x \in [0,1]^d$. Then h_j is supported on $[(j-1)/\kappa, j/\kappa] \subseteq [0,1/3] \times [0,1]^{d-1}$ and $\int h_j^2 = 1$. For $\eta \in \{\pm 1\}^{[\kappa/3]\times[\kappa]^{d-1}}$ and $\rho > 0$ let

$$f_\eta(x) = f_0 + \rho \sum_{j \in [\kappa/3]\times[\kappa]^{d-1}} \eta_j h_j(x).$$

Then f_η is positive provided that $\epsilon^2 \ge \rho\kappa^{d/2}|h|_\infty \asymp \rho\kappa^{d/2}$. Further, $\|f_\eta\|_{\mathcal{C}^\beta}$ is of constant order provided $\rho\kappa^{d/2+\beta} \lesssim 1$. Under these assumptions $f_\eta \in \mathcal{P}_{\mathsf{H}}$. Note that the Hellinger distance between f_η and f_0 is

$$\begin{aligned}
H^2(f_0, f_\eta) &= \sum_{j \in [\kappa/3]\times[\kappa]^{d-1}} \int_{[\frac{j-1}{\kappa}, \frac{j}{\kappa}]} \left(\sqrt{f_0(x)} - \sqrt{f_\eta(x)}\right)^2 \mathrm{d}x \\
&= \sum_{j \in [\kappa/3]\times[\kappa]^{d-1}} \int_{[\frac{j-1}{\kappa}, \frac{j}{\kappa}]} \frac{\rho^2 h_j^2(x)}{(\sqrt{f_0(x)} + \sqrt{f_\eta(x)})^2} \mathrm{d}x \\
&\ge \sum_{j \in [\kappa/3]\times[\kappa]^{d-1}} \int_{[\frac{j-1}{\kappa}, \frac{j}{\kappa}]} \frac{\rho^2 h_j^2(x)}{4\epsilon^2} \mathrm{d}x \\
&\gtrsim \frac{\rho^2\kappa^d}{\epsilon^2}.
\end{aligned}$$

Suppose we draw η uniformly at random. Via Ingster's trick we compute

$$\begin{aligned}
\chi^2(\mathbb{E}_\eta f_\eta^{\otimes n} \| f_0^{\otimes n}) + 1 &= \int \mathbb{E}_{\eta\eta'} \prod_{i=1}^n \frac{f_\eta(x_i) f_{\eta'}(x_i)}{f_0(x_i)} \mathrm{d}x_1 \dots \mathrm{d}x_n \\
&= \mathbb{E}_{\eta\eta'} \left(\int \frac{f_\eta(x) f_{\eta'}(x)}{f_0(x)} \mathrm{d}x \right)^n.
\end{aligned}$$

Looking at the integral term on the inside we get

$$\begin{aligned}\int \frac{f_\eta(x) f_{\eta'}(x)}{f_0(x)} \mathrm{d}x &= \int \frac{\left(f_0(x) + \rho \sum_{j \in [\kappa/3] \times [\kappa]^{d-1}} \eta_j h_j(x)\right)\left(f_0(x) + \rho \sum_{j \in [\kappa/3] \times [\kappa]^{d-1}} \eta'_j h_j(x)\right)}{f_0(x)} \mathrm{d}x \\ &= 1 + \rho \sum_j (\eta_j + \eta'_j) \int h_j(x) \mathrm{d}x + \rho^2 \sum_j \eta_j \eta'_j \int \frac{h_j(x)^2}{f_0(x)} \mathrm{d}x \\ &= 1 + \frac{\rho^2}{\epsilon^2} \sum_j \eta_j \eta'_j \int h_j(x)^2 \mathrm{d}x \\ &= 1 + \frac{\rho^2}{\epsilon^2} \langle \eta, \eta' \rangle,\end{aligned}$$

where we've used that h_j and $h_{j'}$ have disjoint support unless $j = j'$, $\int h_j = 0$, $\int h_j^2 = 1$, and that $f_0(x) = \epsilon^2$ for all x with $x_1 \leq 1/3$. Plugging in, using the inequalities $1 + x \leq \exp(x)$ and $\cosh(x) \leq \exp(x^2)$ we obtain

$$\begin{aligned}\chi^2(\mathbb{E}_\eta f_\eta^{\otimes n} \| f_0^{\otimes n}) + 1 &\leq \mathbb{E}_{\eta\eta'}(1 + \frac{\rho^2}{\epsilon^2}\langle \eta, \eta' \rangle)^n \\ &\leq \mathbb{E}_{\eta\eta'} \exp(\frac{\rho^2 n}{\epsilon^2}\langle \eta, \eta' \rangle) \\ &= \cosh(\frac{\rho^2 n}{\epsilon^2})^{\kappa^d/3} \\ &\leq \exp(\frac{\rho^4 n^2 \kappa^d}{3\epsilon^4}).\end{aligned}$$

Choosing $\kappa = \epsilon^{-2/\beta}$ and $\rho = \epsilon^{(2\beta+d)/\beta}$ we see that goodness-of-fit testing of f_0 is impossible unless

$$n \gtrsim \frac{\epsilon^2}{\rho^2 \kappa^{d/2}} = \epsilon^{-\frac{2\beta+d}{\beta}}.$$

□

A.4 Auxiliary technical results

A.4.1 Proof of Lemma 2.2.1

Proof. We prove the upper bound first. Let $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$ be arbitrary. Then by Lemma A.2.2,

$$\begin{aligned}\inf_\psi \max_{i=0,1} \mathbb{P}_i^{\otimes m}(\psi \neq i) &\leq \inf_\psi \left(\mathbb{P}_0^{\otimes m}(\psi = 1) + \mathbb{P}_1^{\otimes m}(\psi = 0)\right) \\ &= 1 - \mathsf{TV}(\mathbb{P}_0^{\otimes m}, \mathbb{P}_1^{\otimes m}) \\ &\leq 1 - \frac{1}{2}\mathsf{H}^2(\mathbb{P}_0^{\otimes m}, \mathbb{P}_1^{\otimes m}) =: (\dagger).\end{aligned}$$

By tensorization of the Hellinger affinity, we have

$$\mathsf{H}^2(\mathbb{P}_0^{\otimes m}, \mathbb{P}_1^{\otimes m}) = 2 - 2\left(1 - \frac{1}{2}\mathsf{H}^2(\mathbb{P}_0, \mathbb{P}_1)\right)^m. \tag{A.4.1}$$

Plugging in, along with $1 + x \le e^x$ gives

$$(\dagger) \le \exp(-\frac{m}{2}\mathsf{H}^2\left(\mathbb{P}_0^{\otimes m}, \mathbb{P}_1^{\otimes m}\right)).$$

Taking $m > 2\log(3)/\mathsf{H}^2(\mathbb{P}_0, \mathbb{P}_1)$ shows the existence of a successful test. Let us turn to the lower bound. Using Lemma A.2.2 we have

$$\begin{aligned}\inf_\psi \max_{i=0,1} \mathbb{P}_i^{\otimes m}(\psi \neq i) &\ge \frac{1}{2}\left(1 - \mathsf{TV}(\mathbb{P}_0^{\otimes m}, \mathbb{P}_1^{\otimes m})\right) \\ &\ge \frac{1}{2}\left(1 - \mathsf{H}(\mathbb{P}_0^{\otimes m}, \mathbb{P}_1^{\otimes m})\right).\end{aligned}$$

Note that it is enough to restrict the maximization in Lemma 2.2.1 to $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$ with $\mathsf{H}^2(\mathbb{P}_0, \mathbb{P}_1) < 1$. Now, by (A.4.1) and the inequalities $e^{-2x} \le 1 - x$ valid for all $x \in [0, 1/2]$ and $1 - x \le e^{-x}$ valid for all $x \in \mathbb{R}$, we obtain

$$\begin{aligned}\mathsf{H}^2(\mathbb{P}_0^{\otimes m}, \mathbb{P}_1^{\otimes m}) &= 2 - 2\left(1 - \frac{1}{2}\mathsf{H}^2(\mathbb{P}_0, \mathbb{P}_1)\right)^m \\ &\le 2 - 2\exp(-m\mathsf{H}^2(\mathbb{P}_0, \mathbb{P}_1)) \\ &\le 2m\mathsf{H}^2(\mathbb{P}_0, \mathbb{P}_1).\end{aligned}$$

Taking $m = 1/(18\mathsf{H}^2(\mathbb{P}_0, \mathbb{P}_1))$ concludes the proof via Lemma A.2.1. □

A.4.2 Proof of Lemma 2.3.5

Proof. By standard inequalities between divergences (see e.g. Lemma A.2.2), omitting the argument (μ_θ, μ_0) for simplicity we have

$$\mathsf{TV} \le \mathsf{H} \le \sqrt{\mathsf{KL}} \le \sqrt{\chi^2} = \sqrt{\exp(\|\theta\|_2^2) - 1} \lesssim \|\theta\|_2.$$

For the lower bound we obtain $\mathsf{TV}(\mu_\theta, \mu_0) \ge \min\{1, \|\theta\|_2/200\} \gtrsim \|\theta\|_2$ by [61, Theorem 1.2]. □

A.4.3 Proof of Proposition 2.4.3

Let us write $a_+ =: a \vee 0$ for both functions and real numbers. We start with some known results of approximation theory.

Definition 10. *For $f : [0,1]^d \to \mathbb{R}$ define the modulus of continuity as*

$$\omega(\delta; f) = \sup_{\|x-y\|_2 \le \delta} |f(x) - f(y)|.$$

Lemma A.4.1. *For any real-valued function f and $\delta \geq 0$,*

$$\omega(\delta; \sqrt{f_+}) \leq \omega(\delta; f)^{1/2}.$$

Proof. Follows from the inequality $|\sqrt{a_+} - \sqrt{b_+}|^2 \leq |a - b|$ valid for all $a, b \in \mathbb{R}$. □

Lemma A.4.2. *Let $f : [0,1]^d \to \mathbb{R}$ be β-smooth for $\beta \in (0,1]$. Then*

$$\omega(\delta; f) \leq c\, \delta^\beta$$

for a constant c depending only on $\|f\|_{\mathcal{C}^\beta}$.

Proof. Follows by the definition of Hölder continuity. □

Lemma A.4.3 ([156, Theorem 4])**.** *For any continuous function $f : [0,1]^d \to \mathbb{R}$ the best polynomial approximation p_n of degree n satisfies*

$$\|p_n - f\|_\infty \leq c\, \omega\left(\frac{d^{3/2}}{n}; f\right)$$

for a universal constant $c > 0$.

Definition 11. *Given a function $f : [0,1]^d \to \mathbb{R}$, $\ell \geq 1$ and $j \in [\ell]^d$, let $\pi_{j,\ell} f : [0,1]^d \to \mathbb{R}$ denote the function*

$$\pi_{j,\ell} f(x) =: f\left(\frac{x + j - 1}{\ell}\right).$$

In other words, $\pi_{j,\ell} f$ is equal to f zoomed in on the j'th bin of the regular grid $\mathcal{G}_\ell$.

Recall that here P_ℓ denotes the L^2 projector onto the space of functions piecewise constant on the bins of $\mathcal{G}_\ell$. We are ready for the proof of Proposition 2.4.3.

Proof. Let $\kappa \geq r \geq 1$ whose values we specify later. We treat the parameters $\beta, d, \|f\|_{\mathcal{C}^\beta}, \|g\|_{\mathcal{C}^\beta}$ as constants in our analysis. Let $u_f : [0,1]^d \to \mathbb{R}$ denote the (piecewise polynomial) function that is equal to the best polynomial approximation of $\sqrt{f}$ on each bin of $\mathcal{G}_{\kappa/r}$ with maximum degree α. By Lemmas A.4.1 and A.4.2 for any $\ell \geq 1$ and $j \in [\ell]^d$

$$\omega(\delta; \pi_{j,\ell} \sqrt{f}) \leq \omega(\delta/\ell; \sqrt{f}) \lesssim (\delta/\ell)^{\beta/2}, \tag{A.4.2}$$

so that by Lemma A.4.3

$$\begin{aligned} |u_f - \sqrt{f}|_\infty &= \sup_{j \in [\kappa/r]^d} |\pi_{j,\kappa/r}(u_f - \sqrt{f})|_\infty \\ &\lesssim \sup_{j \in [\kappa/r]^d} \omega(d^{3/2}/\alpha; \pi_{j,\kappa/r} \sqrt{f}) \\ &\lesssim (\alpha\kappa/r)^{-\beta/2}. \end{aligned}$$

Regarding r as a constant independent of κ, α can be chosen large enough independently of κ such that $|u_f - \sqrt{f}|_\infty \le c_1\kappa^{-\beta/2}$ for c_1 arbitrarily small. Define u_g analogously to u_f. We have the inequalities

$$\begin{aligned}\mathsf{H}(f,g) &= \|\sqrt{f} - \sqrt{g}\|_2 \\ &\le \|\sqrt{f} - u_f\|_2 + \|u_f - u_g\|_2 + \|u_g - \sqrt{g}\|_2 \\ &\le 2c_1\kappa^{-\beta/2} + \|u_f - u_g\|_2.\end{aligned}$$

We can write

$$\|u_f - u_g\|_2^2 = \frac{1}{(\kappa/r)^d}\sum_{j\in[\kappa/r]^d}\|\pi_{j,\kappa/r}(u_f - u_g)\|_2^2$$

Now, by [10, Lemma 7.4] we can take r large enough (depending only on $\beta, d, \|f\|_{\mathcal{C}^\beta}, \|g\|_{\mathcal{C}^\beta}$) such that

$$\|\pi_{j,\kappa/r}(u_f - u_g)\|_2 \le c_2\|P_r\pi_{j,\kappa/r}(u_f - u_g)\|_2$$

where the implied constant depends on the same parameters as r. Thus, we get

$$\begin{aligned}\mathsf{H}^2(f,g) &\le 8c_1^2\kappa^{-\beta} + \frac{2c_2^2}{(\kappa/r)^d}\sum_{j\in[\kappa/r]^d}\|P_r\pi_{j,\kappa/r}(u_f - u_g)\|_2^2 \\ &\le 8c_1^2\kappa^{-\beta} + \frac{6c_2^2}{(\kappa/r)^d}\sum_{j\in[\kappa/r]^d}\left(\|P_r\pi_{j,\kappa/r}u_f - \sqrt{P_r\pi_{j,\kappa/r}f}\|_2^2 + \|P_r\pi_{j,\kappa/r}u_g - \sqrt{P_r\pi_{j,\kappa/r}g}\|_2^2\right) \\ &\quad + 6c_2^2\mathsf{H}^2(P_\kappa f, P_\kappa f),\end{aligned}$$

where c_1, c_2 depend only on the unimportant parameters, and c_1 can be taken arbitrarily small compared to c_2. We also used the fact that $P_r\pi_{j,\kappa/r} = \pi_{j,\kappa/r}P_\kappa$. Looking at the terms separately, we have

$$\begin{aligned}\|P_r\pi_{j,\kappa/r}u_f - \sqrt{P_r\pi_{j,\kappa/r}f}\|_2 &\le \|P_r\pi_{j,\kappa/r}u_f - P_r\sqrt{\pi_{j,\kappa/r}f}\|_2 + \|P_r\sqrt{\pi_{j,\kappa/r}f} - \sqrt{P_r\pi_{j,\kappa/r}f}\|_2 \\ &\le c\kappa^{-\beta/2} + \|P_r\sqrt{\pi_{j,\kappa/r}f} - \sqrt{P_r\pi_{j,\kappa/r}f}\|_2,\end{aligned}$$

since P_r is a contraction by Lemma A.4.4. We can decompose the second term as

$$\begin{aligned}&\|P_r\sqrt{\pi_{j,\kappa/r}f} - \sqrt{P_r\pi_{j,\kappa/r}f}\|_2^2 = \\ &\quad = \sum_{\ell\in[r]^d}\int_{\left[\frac{\ell-1}{r},\frac{\ell}{r}\right]}\left(r^d\int_{\left[\frac{\ell-1}{r},\frac{\ell}{r}\right]}\sqrt{\pi_{j,\kappa/r}f(x)}\mathrm{d}x - \sqrt{r^d\int_{\left[\frac{\ell-1}{r},\frac{\ell}{r}\right]}\pi_{j,\kappa/r}f(x)\mathrm{d}x}\right)^2 = (\dagger).\end{aligned}$$

For $x \in [(\ell-1)/r, \ell/r]$ we always have

$$|\pi_{j,\kappa/r}f(x) - \pi_{j,\kappa/r}f(\ell/r)| \le \omega(\frac{\sqrt{d}}{r}; \pi_{j,\kappa/r}f) \lesssim \left(\frac{\sqrt{d}/r}{\kappa/r}\right)^\beta \lesssim \kappa^{-\beta}.$$

Using the inequality $\sqrt{a+b} - \sqrt{(a-b)_+} \le 2\sqrt{b}$ valid for all $a, b \ge 0$, we can bound $(\dagger)$ by $\kappa^{-\beta}$ up to constant and the result follows. □

A.4.4 Proof of Proposition A.1.1

For $f \in L^2(\mu)$ write $f_i = \langle f\phi_i\rangle$ and $f_{ii'} = \langle f\phi_i\phi_{i'}\rangle$, assuming that the quantities involved are well-defined. We record some useful identities related to P_r that will be instrumental in our proof of Proposition A.1.1.

Lemma A.4.4. *P_r is self-adjoint and has operator norm*

$$\|P_r\| =: \sup_{f\in L^2(\mu):\|f\|_2\le 1} \|P_r(f)\|_2 \le 1.$$

Suppose that $f, g, h, t \in L^2(\mu)$ and that each quantity below is finite. Then

$$\sum_{ii'} f_i g_{i'} h_{ii'} = \langle hP_r(f)P_r(g)\rangle,$$
$$\sum_{ii'} f_i g_i h_{i'} t_{i'} = \langle fP_r(g)\rangle\langle hP_r(t)\rangle$$
$$\sum_{ii'} f_{ii'} g_{ii'} = \sum_i \langle f\phi_i P_r(g\phi_i)\rangle,$$

where the summation is over $i, i' \in [r]$.

Proof. Let $P_r^\perp$ be the projection onto the orthogonal complement of $\mathrm{span}(\{\phi_1, \dots, \phi_r\})$. Then for any $f, g \in L^2(\mu)$ we have

$$\langle fP_r(g)\rangle = \langle (P_r(f) + P_r^\perp(f))P_r(g)\rangle = \langle P_r(f)P_r(g)\rangle = \langle P_r(f)g\rangle,$$

where the last equality is by symmetry. We also have

$$\|P_r(f)\|_2^2 \le \|P_r(f)\|_2^2 + \|P_r^\perp(f)\|_2^2 = \|P_r(f) + P_r^\perp(f)\|^2 = \|f\|_2^2.$$

Let $f, g, h, t \in L^2(\mu)$. Then

$$\sum_{ii'} f_i g_{i'} h_{ii'} = \sum_i f_i \sum_{i'} g_{i'} h_{ii'} = \sum_i f_i \sum_{i'} \langle gP_r(h\phi_i)\rangle = \sum_i f_i \langle P_r(g)h\phi_i\rangle = \langle P_r(f)hP_r(g)\rangle$$
$$\sum_{ii'} f_i g_i h_{i'} t_{i'} = (\sum_i f_i g_i)(\sum_{i'} h_{i'} t_{i'}) = \langle fP_r(g)\rangle\langle hP_r(t)\rangle$$
$$\sum_{ii'} f_{ii'} g_{ii'} = \sum_i \langle f\phi_i \sum_{i'} \langle g\phi_i\phi_{i'}\rangle\phi_{i'}\rangle = \sum_i \langle f\phi_i P_r(g\phi_i)\rangle.$$

□

Proof of Proposition A.1.1. Let us label the different terms of the statistic $T_{\mathsf{LF}}^{-\mathsf{d}}$:

$$\begin{aligned} T_{\mathsf{LF}}^{-\mathsf{d}} &= \sum_{i=1}^r \Bigg\{ \frac{2}{n^2}\sum_{j<j'}^n \phi_i(X_j)\phi_i(X_{j'}) - \frac{2}{n^2}\sum_{j<j'}^n \phi_i(Y_j)\phi_i(Y_{j'}) \\ &\qquad - \frac{2}{nm}\sum_{j=1}^n\sum_{u=1}^m \phi_i(X_j)\phi_i(Z_u) + \frac{2}{nm}\sum_{j=1}^n\sum_{u=1}^m \phi_i(Y_j)\phi_i(Z_u)\Bigg\} \\ &= \frac{2}{n^2}\mathsf{I} - \frac{2}{n^2}\mathsf{II} - \frac{2}{nm}\mathsf{III} + \frac{2}{nm}\mathsf{IV}. \end{aligned}$$

Recall that $X, Y, Z \sim f^{\otimes n}, g^{\otimes n}, h^{\otimes m}$ respectively. A straightforward computation yields

$$\mathbb{E}T_{\mathsf{LF}} = \|P_r(f-h)\|_2^2 - \|P_r(g-h)\|_2^2 - \frac{1}{n}\big(\|P_r(f)\|_2^2 - \|P_r(g)\|_2^2\big).$$

We decompose the variance as

$$\begin{aligned}\operatorname{var}(T_{\mathsf{LF}}) = \frac{4}{n^4}\operatorname{var}(\mathsf{I}) + \frac{4}{n^4}\operatorname{var}(\mathsf{II}) + \frac{4}{n^2m^2}\operatorname{var}(\mathsf{III}) + \frac{4}{n^2m^2}\operatorname{var}(\mathsf{IV}) \\ - \frac{8}{n^3m}\operatorname{Cov}(\mathsf{I},\mathsf{III}) - \frac{8}{n^3m}\operatorname{Cov}(\mathsf{II},\mathsf{IV}) - \frac{8}{n^2m^2}\operatorname{Cov}(\mathsf{III},\mathsf{IV}),\end{aligned}$$

where we used independence of the pairs $(\mathsf{I},\mathsf{II}), (\mathsf{I},\mathsf{IV}), (\mathsf{II},\mathsf{III})$. Expanding the variances we obtain

$$\begin{aligned}\operatorname{var}(\mathsf{I}) &= \sum_{ii'}\left(\binom{n}{2}(f_{ii'}^2 - f_i^2 f_{i'}^2) + (\binom{n}{2}^2 - \binom{n}{2} - \binom{4}{2}\binom{n}{4})(f_i f_{i'} f_{ii'} - f_i^2 f_{i'}^2)\right) \\ \operatorname{var}(\mathsf{II}) &= \sum_{ii'}\left(\binom{n}{2}(g_{ii'}^2 - g_i^2 g_{i'}^2) + (\binom{n}{2}^2 - \binom{n}{2} - \binom{4}{2}\binom{n}{4})(g_i g_{i'} g_{ii'} - g_i^2 g_{i'}^2)\right) \\ \operatorname{var}(\mathsf{III}) &= \sum_{ii'}\Big(nm(f_{ii'}h_{ii'} - f_i f_{i'} h_i h_{i'}) + nm(m-1)(f_{ii'}h_i h_{i'} - f_i f_{i'} h_i h_{i'}) + \\ &\qquad + mn(n-1)(f_i f_{i'} h_{ii'} - f_i f_{i'} h_i h_{i'})\Big) \\ \operatorname{var}(\mathsf{IV}) &= \sum_{ii'}\Big(nm(h_{ii'}g_{ii'} - h_i h_{i'} g_i g_{i'}) + mn(n-1)(h_{ii'}g_i g_{i'} - h_i h_{i'} g_i g_{i'}) \\ &\qquad + nm(m-1)(g_{ii'}h_i h_{i'} - h_i h_{i'} g_i g_{i'})\Big).\end{aligned}$$

For the covariance terms we obtain

$$\begin{aligned}\operatorname{Cov}(\mathsf{I},\mathsf{III}) &= \sum_{ii'} 2m\binom{n}{2}(f_{ii'}f_i h_{i'} - f_i^2 f_{i'} h_{i'}) \\ \operatorname{Cov}(\mathsf{II},\mathsf{IV}) &= \sum_{ii'} 2m\binom{n}{2}(g_{ii'}g_i h_{i'} - g_i^2 g_{i'} h_{i'}) \\ \operatorname{Cov}(\mathsf{III},\mathsf{IV}) &= \sum_{ii'} mn^2(h_{ii'}f_i g_{i'} - f_i g_{i'} h_i h_{i'}).\end{aligned}$$

We can now start collecting the terms, applying the calculation rules from Lemma A.4.4 repeatedly. Note that $\binom{n}{2}^2 - \binom{n}{2} - \binom{4}{2}\binom{n}{4} = n^3 - 3n^2 + 2n$, and by inspection we can conclude that $1/n, 1/m, 1/nm, 1/n^2$ and $1/n^3$ are the only terms with nonzero coefficients. We look at

each of them one-by-one:

$$\begin{aligned}
\text{Coef}\left(\frac{1}{n}\right) &= \sum_{ii'}^{r}\Big(\underbrace{4(f_if_{i'}f_{ii'}-f_i^2f_{i'}^2)}_{\text{var}(I)}+\underbrace{4(g_ig_{i'}g_{ii'}-g_i^2g_{i'}^2)}_{\text{var}(II)}+\underbrace{4(h_ih_{i'}f_{ii'}-f_if_{i'}h_ih_{i'})}_{\text{var}(III)}+\\
&\qquad \underbrace{4(g_{ii'}h_ih_{i'}-h_ih_{i'}g_ig_{i'})}_{\text{var}(IV)}-\underbrace{8(f_{ii'}f_ih_{i'}-f_i^2f_{i'}h_{i'})}_{\text{Cov}(I,III)}-\underbrace{8(g_{ii'}g_ih_{i'}-g_i^2g_{i'}h_{i'})}_{\text{Cov}(II,IV)}\Big)\\
&= 4\langle fP_r(f)^2\rangle-4\langle fP_r(f)\rangle^2+4\langle gP_r(g)^2\rangle-4\langle gP_r(g)\rangle^2+4\langle fP_r(h)^2\rangle-4\langle fP_r(h)\rangle^2\\
&\quad +4\langle gP_r(h)^2-4\langle hP_r(g)\rangle^2-8\langle fP_r(f)P_r(h)\rangle+8\langle fP_r(f)\rangle\langle fP_r(h)\rangle\\
&\quad -8\langle gP_r(g)P_r(h)\rangle+8\langle gP_r(g)\rangle\langle gP_r(h)\rangle\\
&= 4\langle f(P_r(f-h))^2\rangle+4\langle g(P_r(g-h))^2\rangle-4\langle P_r(f-h)\rangle^2-4\langle P_r(g-h)\rangle^2\\
&\le 4A_{ffh}+4A_{ggh},
\end{aligned}$$

recalling the definition $A_{uvt}=\langle u\big[P_r(v-t)\big]^2\rangle$ for $u,v,t\in L^2(\mu)$. Similarly, we get

$$\begin{aligned}
\text{Coef}\left(\frac{1}{m}\right) &= \sum_{ii'}^{r}\Big(\underbrace{4(h_{ii'}f_if_{i'}-f_if_{i'}h_ih_{i'})}_{\text{var}(III)}+\underbrace{4(h_{ii'}g_ig_{i'}-h_ih_{i'}g_ig_{i'})}_{\text{var}(IV)}-\underbrace{8(h_{ii'}f_ig_{i'}-f_ih_ih_{i'}g_{i'})}_{\text{Cov}(III,IV)}\\
&= 4\langle h(P_r(f-g))^2\rangle-4\langle hP_r(f-g)\rangle^2\\
&\le 4A_{hfg}.
\end{aligned}$$

For the lower order terms we obtain

$$\begin{aligned}
\text{Coef}\left(\frac{1}{nm}\right) &= \sum_{ii'}^{r}\Big(\underbrace{4(f_{ii'}h_{ii'}-f_if_{i'}h_ih_{i'})-4(f_{ii'}h_ih_{i'}-f_if_{i'}h_ih_{i'})-4(f_if_{i'}h_{ii'}-f_if_{i'}h_ih_{i'})}_{\text{var}(III)}\\
&\qquad +\underbrace{4(h_{ii'}g_{ii'}-h_ih_{i'}g_ig_{i'})-4(h_{ii'}g_ig_{i'}-h_ih_{i'}g_ig_{i'})-4(g_{ii'}h_ih_{i'}-h_ih_{i'}g_ig_{i'})}_{\text{var}(IV)}\Big)\\
&= 4B_{fh}-4\langle fP_r(h)\rangle^2-4\langle fP_r(h)^2\rangle+4\langle fP_r(h)\rangle^2\\
&\quad -4\langle hP_r(f)^2\rangle+4\langle fP_r(h)\rangle^2+4B_{gh}-4\langle gP_r(h)\rangle^2\\
&\quad -4\langle hP_r(g)^2\rangle+4\langle gP_r(h)\rangle^2-4\langle gP_r(h)^2\rangle+4\langle gP_r(h)\rangle^2\\
&\le 4\langle fP_r(h)\rangle^2+4\langle gP_r(h)\rangle^2+4B_{fh}+4B_{gh}\\
&\lesssim |B_{fh}|+|B_{gh}|+\|f+g+h\|_2^4
\end{aligned}$$

where we recall the definition $B_{uv}=\sum_i\langle u\phi_iP_r(v\phi_i)\rangle$ for $u,v\in L^2(\mu)$ and apply the Cauchy-

Schwarz inequality. Next, we look at the coefficient of $1/n^2$ and find

$$\begin{aligned}
\mathrm{Coef}\left(\frac{1}{n^2}\right) &= \sum_{ii'}\Big(\underbrace{2(f_{ii'}^2 - f_i^2 f_{i'}^2) - 12(f_{ii'}f_i f_{i'} - f_i^2 f_{i'}^2)}_{\mathrm{var(I)}} + \underbrace{2(g_{ii'}^2 - g_i^2 g_{i'}^2) - 12(g_{ii'}g_i g_{i'} - g_i^2 g_{i'}^2)}_{\mathrm{var(II)}} \\
&\qquad + \underbrace{8(f_{ii'}f_i h_{i'} - f_i^2 f_{i'}h_{i'})}_{\mathrm{Cov(I,III)}} + \underbrace{8(g_{ii'}g_i h_{i'} - g_i^2 g_{i'}h_{i'})}_{\mathrm{Cov(II,IV)}}\Big) \\
&= 2B_{ff} - 2\langle fP_r(f)\rangle^2 - 12\langle fP_r(f)^2\rangle + 12\langle fP_r(f)\rangle^2 \\
&\qquad + 2B_{gg} - 2\langle gP_r(g)\rangle^2 - 12\langle gP_r(g)^2\rangle + 12\langle gP_r(g)\rangle^2 \\
&\qquad + 8\langle fP_r(f)P_r(h)\rangle - 8\langle fP_r(f)\rangle\langle fP_r(h)\rangle + 8\langle gP_r(g)P_r(h)\rangle - 8\langle gP_r(g)\rangle\langle gP_r(h)\rangle \\
&\le 2B_{ff} + 2B_{gg} + 8\langle fP_r(f)P_r(h-f)\rangle + 8\langle gP_r(g)P_r(h-g)\rangle + 40\|f+g+h\|_2^4 \\
&\lesssim |B_{ff}| + |B_{gg}| + \|f+g+h\|_2^4 + \sqrt{A_{ff0}A_{ffh} + A_{gg0}A_{ggh}}.
\end{aligned}$$

Finally, we look at the coefficient of $1/n^3$:

$$\begin{aligned}
\mathrm{Coef}\left(\frac{1}{n^3}\right) &= \sum_{ii'}\Big(\underbrace{-2(f_{ii'}^2 - f_i^2 f_{i'}^2) + 8(f_{ii'}f_i f_{i'} - f_i^2 f_{i'}^2)}_{\mathrm{Cov(I,III)}}\underbrace{-2(g_{ii'}^2 - g_i^2 g_{i'}^2) + 8(g_{ii'}g_i g_{i'} - g_i^2 g_{i'}^2)}_{\mathrm{Cov(I,III)}}\Big) \\
&= -2B_{ff} + 2\langle fP_r(f)\rangle^2 + 8\langle fP_r(f)^2\rangle - 8\langle fP_r(f)\rangle^2 \\
&\qquad - 2B_{gg} + 2\langle gP_r(g)\rangle^2 + 8\langle gP_r(g)^2\rangle - 8\langle gP_r(g)\rangle^2 \\
&\lesssim |B_{ff}| + |B_{gg}| + \|f+g+h\|_2^4 + A_{ff0} + A_{gg0}.
\end{aligned}$$

□

A.4.5 Proof of Lemma A.2.3

Proof. Expanding via the binomial formula and using the fact that sums of N_j's are binomial random variables, we get

$$\begin{aligned}
\mathbb{E}_N \prod_{j\in k}(a + b(1+c)^{N_j}) &= \mathbb{E}\sum_{\ell=0}^{k}\binom{k}{\ell} b^\ell (1+c)^{\mathrm{Bin}(n,\ell/k)} a^{k-\ell} \\
&= \sum_{\ell=0}^{k}\binom{k}{\ell} b^\ell (1+\frac{c\ell}{k})^n a^{k-\ell} \\
&\le (a + be^{cn/k})^k,
\end{aligned}$$

where we used $1 + x \le e^x$ for all $x \in \mathbb{R}$. □

Appendix B

Appendix of "Kernel-Based Tests for Likelihood-Free Hypothesis Testing"

B.1 Notation

We use $A \gtrsim B, A \lesssim B, A \asymp B$ to denote $A = \Omega(B), B = \Omega(A)$ and $A = \Theta(B)$ respectively, where the hidden constants depend on untracked parameters multiplicatively.[1]

We write $\mathsf{TV}, \mathsf{KL}, \chi^2$ for total-variation, KL-divergence and χ^2-divergence, respectively. We write $D(P_{Y|X}\|Q_{Y|X}|P_X) = \mathbb{E}_{X\sim P_X} D(P_{Y|X}\|Q_{Y|X})$ as the *conditional divergence* for any probability measures P, Q on two variables X, Y and divergence $D \in \{\mathsf{TV}, \mathsf{KL}, \chi^2\}$.

We write ℓ^p for the usual ℓ^p sequence space and L^p for the usual L^p space with respect to the Lebesgue measure. Both the ℓ^p norm and the L^p norm are written as $\|\cdot\|_p$ if no ambiguity arises.

For real numbers $a, b \in \mathbb{R}$ we also write $\max\{a, b\}$ as $a \vee b$ and $\min\{a, b\}$ as $a \wedge b$.

We use $\vec{1}_d$ to denote an d-dimensional all 1's vector.

For an integer $k \in \mathbb{Z}^+$, we write $[k]$ as a short notation for the set $\{1, 2, \ldots, k\}$.

In the proofs of Theorem 3.3.1 and Theorem 3.3.2, we use $\overset{!}{=}$ for an equality that we are trying to prove.

B.2 Applications of Theorem 3.3.1

Usually, minimax rates of testing are proven under separation assumptions using more traditional measures of distance such as L^p, where $p \in [1, \infty]$. In this section we show one example of how Theorem 3.3.1 can be used to recover known results, and also obtain some novel results under L^2-separation and L^1-separation.

[1] For example, the first equation in (B.2.1) means that there exists a constant c independent of $\alpha, k, \epsilon, \delta, R$, such that $\min\{m, n\} \geq c\frac{\log(1/\alpha)(1+R)^2}{k\epsilon^2\delta^2}$.

B.2.1 Bounded Discrete Distributions Under L^2/L^1-Separation

Sample Complexity Upper Bounds Let $\mathcal{P}_{\mathsf{Db}}(k, C)$ be the set of all discrete distributions P supported on $[k] = \{1, 2, \ldots, k\}$ satisfying $\max_{1 \le i \le k} p(i) \le C/k$, where p is the probability mass function of P (here $\sum_{i=1}^k p(k) = 1$). For distributions P_X, P_Y, P_Z we shall write p_X, p_Y, p_Z as their probability mass functions, respectively.

Let us apply Theorem 3.3.1 with underlying space $\mathcal{X} = [k]$ and measure $\mu = \frac{1}{k}\sum_{i=1}^k \delta_i$. Take the kernel $K(x, y) = \mathbb{1}\{x = y\} = \sum_{i=1}^k \mathbb{1}\{x = y = i\}$, and note that for any two distributions P_X, P_Y we have

$$\mathrm{MMD}^2(P_X, P_Y) = \mathbb{E}\Big[K(X, X') + K(Y, Y') - 2K(X, Y)\Big] = \sum_i |p_X(i) - p_Y(i)|^2$$

where $(X, X', Y, Y') \sim P_X^{\otimes 2} \otimes P_Y^{\otimes 2}$. So the corresponding MMD is the ℓ^2-distance on probability mass functions. Note also that $K = \sum_{i=1}^k \frac{1}{k}\Big(\sqrt{k}\mathbb{1}\{x = i\}\Big)\Big(\sqrt{k}\mathbb{1}\{y = i\}\Big)$, where $\Big\{\sqrt{k}\mathbb{1}\{x = i\}\Big\}_{i=1}^k$ forms an orthonormal basis of $L^2(\mu)$. So K has only one nonzero eigenvalue, namely

$$\lambda_1 = \lambda_2 = \ldots = \lambda_k = 1/k,$$

of multiplicity k. Suppose that we observe samples X, Y, Z of size n, n, m from $P_X, P_Y, P_Z \in \mathcal{P}_{\mathsf{Db}}(k, C)$, where $\mathrm{MMD}(P_X, P_Y) = \sqrt{\sum_i |p_X(i) - p_Y(i)|^2} \ge \epsilon$. Plugging into Theorem 3.3.1 shows that:

Proposition B.2.1. *For any two $P_X, P_Y \in \mathcal{P}_{\mathsf{Db}}(k, C)$, if the ℓ^2-distance between p_X, p_Y is at least ϵ, then testing (mLFHT) is possible at total error α using n simulation samples and m real data samples provided that*

$$\begin{aligned} \min\{m, n\} &\gtrsim \frac{C\|\lambda\|_\infty \log(1/\alpha)(1+R)^2}{\delta^2\epsilon^2} \asymp \frac{\log(1/\alpha)(1+R)^2}{k\epsilon^2\delta^2}, \\ \min\{n, \sqrt{mn}\} &\gtrsim \frac{C\|\lambda\|_2 \sqrt{\log(1/\alpha)}}{\epsilon^2\delta} \asymp \frac{\sqrt{\log(1/\alpha)}}{\sqrt{k}\epsilon^2\delta}. \end{aligned} \tag{B.2.1}$$

where R is defined as in the assumption (iii) of Section 3.3.1.

We can convert the above results to measure separation with respect to total variation (recall $\mathsf{TV}(p, q) = \frac{1}{2}\sum_i |p(i) - q(i)| = \frac{1}{2}\|p - q\|_1$) using the AM-QM inequality $\|p_X - p_Y\|_1 \le \sqrt{k}\|p_X - p_Y\|_2$. Then, taking $R \asymp \alpha \asymp \delta = \Theta(1)$ recovers the minimax optimal results of [123, 124, 78], for LFHT over the class $\mathcal{P}_{\mathsf{Db}}$. Note that analogous results for two-sample testing follow from the above using the reduction presented in Section 3.3.4.

Sample Complexity Lower Bounds Recall the definition of $J_\epsilon^\star$ and note that $\|\lambda\|_{2,J}^2 = \frac{\min(J-1,k)}{k^2}$ for all $J \ge 2$. By Corollary 3.3.3 we see that $J_\epsilon^\star \gtrsim k$ as soon as $\epsilon \lesssim 1/k$. Thus, for $\epsilon \lesssim 1/k$ the necessity of

$$m \gtrsim \frac{\log(1/\alpha)}{k\epsilon^2\delta^2},\ n \gtrsim \frac{\sqrt{\log(1/\alpha)}}{\sqrt{k}\epsilon^2} \ \text{ and } \ m + \sqrt{mn} \gtrsim \frac{\sqrt{\log(1/\alpha)}}{\sqrt{k}\epsilon^2\delta} \tag{B.2.2}$$

follows by Theorem 3.3.2. Here it is crucial to note that when $\delta = \Theta(1)$, we have

$$m + \sqrt{mn} \gtrsim \frac{\sqrt{\log(1/\alpha)}}{\sqrt{k}\epsilon^2} \text{ and } n \gtrsim \frac{\sqrt{\log(1/\alpha)}}{\sqrt{k}\epsilon^2} \iff \sqrt{mn} \gtrsim \frac{\sqrt{\log(1/\alpha)}}{\sqrt{k}\epsilon^2} \text{ and } n \gtrsim \frac{\sqrt{\log(1/\alpha)}}{\sqrt{k}\epsilon^2}$$

and hence the upper bound (B.2.1) meets with the lower bound (B.2.2) provided $R \asymp \delta = \Theta(1)$. Once again, setting $R \asymp \delta \asymp \alpha = \Theta(1)$ we the optimal lower bounds recovering the results of [78] (in the regime $\epsilon \lesssim 1/k$). In short we can also recover the following result for LFHT.

Proposition B.2.2 ([78, Theorem 1, adapted])**.** *On the class $P_{\mathsf{Db}}(k, C)$, using n simulation samples and m real data samples, if*

$$n \gtrsim \frac{1}{\sqrt{k}\epsilon^2}, \quad m \gtrsim \frac{1}{k\epsilon^2}, \quad \sqrt{mn} \gtrsim \frac{1}{\sqrt{k}\epsilon^2}, \tag{B.2.3}$$

then for any two distributions $P_X, P_Y \in \mathcal{P}_{\mathsf{Db}}(k, C)$ with $\|p_X - p_Y\|_2 \geq \epsilon$, testing (LFHT) is possible with a total error of 1%. Conversely, to ensure the existence of a procedure that can test (LFHT) with a total error of 1% for any $P_X, P_Y \in \mathcal{P}_{\mathsf{Db}}(k, C)$ with $\|p_X - p_Y\|_2 \geq \epsilon$, the number of observations (n, m) must satisfy

$$n \gtrsim \frac{1}{\sqrt{k}\epsilon^2}, \quad m \gtrsim \frac{1}{k\epsilon^2}, \quad \sqrt{mn} \gtrsim \frac{1}{\sqrt{k}\epsilon^2}. \tag{B.2.4}$$

The implied constants in (B.2.3) and (B.2.4) do not depend on k and ϵ, but may differ.

B.2.2 β-Hölder Smooth Densities on $[0, 1]^d$ Under L^2/L^1-Separation

Sample Complexity Upper Bounds Let $\mathcal{P}_{\mathsf{H}}(\beta, d, C)$ be the set of all distributions on $[0, 1]^d$ with β-Hölder smooth Lebesgue-density p satisfying $\|p\|_{\mathcal{C}^\beta} \leq C$ for some constant $C > 1$, where

$$\|p\|_{\mathcal{C}^\beta} := \max_{0 \leq |\alpha| \leq \lceil \beta - 1 \rceil} \|f^{(\alpha)}\|_\infty + \sup_{x \neq y \in [0,1]^d, |\alpha| = \lceil \beta - 1 \rceil} \frac{|f^{(\alpha)}(x) - f^{(\alpha)}(y)|}{\|x - y\|_2^{\beta - \lceil \beta - 1 \rceil}},$$

where $\lceil \beta - 1 \rceil$ is the largest integer strictly smaller than β and $|\alpha| = \sum_i \alpha_i$ is the norm of a multi-index $\alpha \in \mathbb{N}^d$. Abusing notation, we also use $\mathcal{P}_{\mathsf{H}}(\beta, d, C)$ to denote the set of all corresponding density functions.

We take $K(x, y) = \sum_j \mathbb{1}\{x, y \in B_j\}$, where $\{B_j\}_{j \in [\kappa]^d}$ is the j'th cell of the regular grid of size κ^d on $[0, 1]^d$, i.e., $B_j = [(j - \vec{1}_d)/\kappa, j/\kappa]$ for $j \in [\kappa]^d$. Clearly there are κ^d nonzero eigenvalues, each equal to 1. The following approximation result is due to Ingster [110], see also [10, Lemma 7.2].

Lemma B.2.3. *Let $f, g \in \mathcal{P}_{\mathsf{H}}(\beta, d, C)$ with $\|f - g\|_2 \geq \epsilon$. Then, there exist constants c, c' independent of ϵ such that for any $\kappa \geq c\epsilon^{-1/\beta}$,*

$$\mathrm{MMD}(f, g) \geq c' \|f - g\|_2.$$

Now, suppose that we have samples X, Y, Z of size n, n, m from $P_X, P_Y, P_Z \in \mathcal{P}_{\mathsf{H}}(\beta, d, C)$ with densities p_X, p_Y, p_Z such that $\|p_X - p_Y\|_2 \geq \epsilon$. Then, Theorem 3.3.1 combined with Lemma B.2.3 and the choice $\kappa \asymp \epsilon^{-1/\beta}$ shows that

Proposition B.2.4. *Testing* (mLFHT) *on* $\mathcal{P}_{\mathsf{H}}(\beta, d, C)$ *at total error* α *using* n *simulation and* m *real data samples is possible provided*

$$\min\{m, n\} \gtrsim \frac{C\|\lambda\|_\infty \log(1/\alpha)(1+R)^2}{\delta^2\epsilon^2} \asymp \frac{\log(1/\alpha)(1+R)^2}{\delta^2\epsilon^2},$$
$$\min\{n, \sqrt{nm}\} \gtrsim \frac{C\|\lambda\|_2 \sqrt{\log(1/\alpha)}}{\epsilon^2\delta} \asymp \frac{\sqrt{\log(1/\alpha)}}{\epsilon^{(2\beta+d/2)/\beta}\delta},$$

where ϵ *is an* L^2*-distance lower bound between* P_X, P_Y *and* R *is defined as in the assumption* (*iii*) *of Section 3.3.1.*

Setting $R \asymp \alpha \asymp \delta = \Theta(1)$ recovers the optimal results of [78] for the class $\mathcal{P}_{\mathsf{H}}$. Once again, identical results under L^1 separation follow from Jensen's inequality $\|\cdot\|_{L^1([0,1]^d)} \leq \|\cdot\|_{L^2([0,1]^d)}$. Note that analogous results for two-sample testing follow from the above using the reduction presented in Section 3.3.4.

Sample Complexity Lower Bounds The kernel defined in the previous paragraph is not suitable for constructing lower bounds over the class $\mathcal{P}_{\mathsf{H}}$ because its eigenfunctions do not necessarily lie in $\mathcal{P}_{\mathsf{H}}$. It would be possible to consider a different kernel that is more adapted to this problem/class but we do not pursue this here.

B.2.3 $(\beta, 2)$-Sobolev Smooth Densities on $\mathbb{R}^d$ Under L^2-Separation

Sample Complexity Upper Bounds Let $\mathcal{P}_{\mathsf{S}}(\beta, d, C)$ be the class of distributions that are supported on $\mathbb{R}^d$ and whose Lebesgue density p satisfies $\|p\|_{\beta,2} \leq C$, where

$$\|p\|_{\beta,2} := \left\|(1 + \|\cdot\|)^\beta \mathcal{F}[p]\right\|_2 \tag{B.2.5}$$

and $\mathcal{F}$ denotes the Fourier transform. Again, abusing notation, we write $\mathcal{P}_{\mathsf{S}}(\beta, d, C)$ both as the set of distributions and the set of density functions.

We take the Gaussian kernel $G_\sigma(x, y) = \sigma^{-d}\exp(-\|x - y\|_2^2/\sigma^2)$ on $\mathcal{X} = \mathbb{R}^d$ with base measure $\mathrm{d}\mu(x) = \exp(-x^2)\mathrm{d}x$. In [139] the authors showed that the two-sample test that thresholds the Gaussian MMD with appropriately chosen variance σ^2 achieves the minimax optimal sample complexity over $\mathcal{P}_{\mathsf{S}}$, when separation is measured by L^2. A key ingredient in their proof is the following inequality.

Lemma B.2.5 ([139, Lemma 5])**.** *Let* $f, g \in \mathcal{P}_{\mathsf{S}}(\beta, d, C)$ *with* $\|f - g\|_2 \geq \epsilon$*. Then, there exist constants* c, c' *independent of* ϵ *such that for any* $\sigma \leq c\,\epsilon^{1/\beta}$*, we have*

$$\mathrm{MMD}(f, g) \geq c'\|f - g\|_2.$$

Now, suppose that we have samples X, Y, Z of sizes n, n, m from $P_X, P_Y, P_Z \in \mathcal{P}_{\mathsf{S}}(\beta, d, C)$ for some constant C with densities p_X, p_Y, p_Y satisfying $\|p_X - p_Y\|_2 \geq \epsilon$.

Note that the heat-semigroup is an L^2-contraction ($\|\lambda\|_\infty \leq 1$) and that

$$\|\lambda\|_2^2 = \int G_\sigma(x,y)^2 \mathrm{d}\mu(x)\mathrm{d}\mu(y) \asymp \sigma^{-d}$$

up to constants depending on the dimension. Theorem 3.3.1 combined with Lemma B.2.5 and a choice $\sigma \asymp \epsilon^{1/\beta}$ yields the following result.

Proposition B.2.6. *Testing* (mLFHT) *over the class* $\mathcal{P}_{\mathsf{S}}$ *with total error* α *is possible provided*

$$\min\{m,n\} \gtrsim \frac{C\|\lambda\|_\infty \log(1/\alpha)(1+R)^2}{\delta^2\epsilon^2} \asymp \frac{\log(1/\alpha)(1+R)^2}{\delta^2\epsilon^2}$$
$$\min\{n,\sqrt{nm}\} \gtrsim \frac{C\|\lambda\|_2\sqrt{\log(1/\alpha)}}{\epsilon^2\delta} \asymp \frac{\sqrt{\log(1/\alpha)}}{\epsilon^{(2\beta+d/2)/\beta}\delta},$$

where ϵ *is the lower bound on the* L^2*-distance between* P_X, P_Y *and* R *is defined as in the assumption* (*iii*) *of Section 3.3.1.*

Taking $R \asymp \delta \asymp \alpha = \Theta(1)$ above, we obtain new results for LFHT and using the reduction from two-sample testing given in Section 3.3.4 we partly recover [139, Theorem 5]. Only partly, because the above requires bounded density with respect to our base measure $\mathrm{d}\mu(x) = \exp(-x^2)\mathrm{d}x$.

Sample Complexity Lower Bounds Note that our lower bound Theorem 3.3.2 doesn't apply because the top eigenfunction of the Gaussian kernel is not constant. Once again, a more careful choice of the base measure (or kernel) might lead to a more suitable argument for the lower bound. We leave such pursuit as open.

B.3 Black-box Boosting of Success Probability

In this section we briefly describe how upper bounds on the minimax sample complexity in the constant error probability regime ($\alpha = \Theta(1)$) can be used to obtain the dependence $\log(1/\alpha)$ in the small error probability regime ($\alpha = o(1)$). We will argue abstractly in a way that applies to the setting of Theorem 3.3.1.

Suppose that from some distributions $P_1, P_2, \ldots, P_k$ we take samples $X^1, X^2, \ldots, X^k$ of size $n_1, n_2, \ldots, n_k$ respectively and are able to decide between two hypotheses H_0 and H_1 (fixed but arbitrary) with total error probability at most $1/3$. Call this test as $\Psi(X^1, \ldots, X^k) \in \{0,1\}$, so that

$$\mathbb{P}(\Psi(X^1,\ldots,X^k) = 0 | H_0) \geq 2/3 \qquad \text{and} \qquad \mathbb{P}(\Psi(X^1,\ldots,X^k) = 1 | H_1) \geq 2/3.$$

Now, to each an error of $o(1)$, instead, we take $18n_1\log(2/\alpha), \ldots, 18n_k\log(2/\alpha)$ observations from P_1 through P_k, and split each sample into $18\log(2/\alpha)$ equal sized batches $\{X^{i,j}\}_{i\in[k],j\in[18\log(2/\alpha)]}$. Here $18\log(2/\alpha)$ is assumed to be an integer without loss of generality. The split samples form $18\log(2/\alpha)$ independent binary random variables

$$A_j := \Psi(X^{1,j}, \ldots, X^{k,j})$$

for $j = 1, 2, \ldots, 18\log(2/\alpha)$. We claim that the majority voting test

$$\Psi_\alpha(\{X^{i,j}\}_{i,j}) = \begin{cases} 1 & \text{if } \bar{A} \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

tests H_0 against H_1 with total probability of error at most α, where

$$\bar{A} := \frac{1}{18\log(2/\alpha)} \sum_{j=1}^{18\log(2/\alpha)} A_j.$$

Indeed, by Hoeffding's inequality, we have

$$\begin{aligned} \mathbb{P}\left(\bar{A} \geq 1/2 \middle| H_0\right) &\leq \alpha/2 \\ \mathbb{P}\left(\bar{A} \leq 1/2 \middle| H_1\right) &\leq \alpha/2. \end{aligned}$$

Therefore, in the remainder of our upper bound proofs, we only focus on achieving a constant probability of error ($\alpha = \Theta(1)$) as the logarithmic dependence follows by the above.

Remark 21. *As mentioned in the discussion succeeding Corollary 3.3.3, we do conjecture the* tight *dependence in the upper bound to be* $\sqrt{\log(\alpha^{-1})}$ *instead of* $\log(\alpha^{-1})$ *shown by this method.*

B.4 Proof of Theorem 3.3.1

B.4.1 Notation and Technical Tools

We use the expansion

$$K(x, y) = \sum_\ell \lambda_\ell e_\ell(x) e_\ell(y)$$

extensively, where $\lambda := (\lambda_1, \lambda_2, \ldots)$ are K's eigenvalues (regarded as an integral operator on $L^2(\mu)$) in non-increasing order and $e_1, e_2, \ldots$ are the corresponding eigenfunctions forming an orthonormal basis for $L^2(\mu)$, and convergence is to be understood in $L^2(\mu)$. We use the notation $\langle \cdot \rangle := \int \cdot \, \mathrm{d}\mu$. For all $u \in L^2(\mu)$ we define

$$u_\ell := \langle u e_\ell \rangle, \quad u_{\ell\ell'} := \langle u e_\ell e_{\ell'} \rangle, \quad \ell = 1, 2, \ldots$$

and consequently $u = \sum_\ell u_\ell e_\ell$. We also define

$$K[u](\cdot) := \int K(t, \cdot) u(t) \mu(\mathrm{d}t) = \sum_\ell \lambda_\ell u_\ell e_\ell(\cdot),$$

where the second equality follows from the orthonormality of $\{e_\ell\}_{\ell=1}^\infty$. Note that the RKHS embedding satisfies $\theta_u := \int K(x, \cdot) u(x) \mathrm{d}\mu(x) = K[u]$. Now, for P_X we write

$$x_\ell := (p_X)_\ell = \langle p_X e_\ell \rangle, \quad x_{\ell\ell'} := (p_X)_{\ell\ell'} = \langle p_X e_\ell e_{\ell'} \rangle, \quad \ell, \ell' = 1, 2, \ldots$$

where p_X is the μ-density of P_X. The similar notations also apply to P_Y, P_Z. The following identities will be very useful in our proofs.

Lemma B.4.1. *For each identity below, let $f, g, h \in L^2(\mu)$ be such that the quantity is well defined. Then,*

$$\|\theta_f\|^2_{\mathcal{H}_K} = \sum_\ell \lambda_\ell f_\ell^2 \tag{B.4.1}$$

$$\mathrm{MMD}^2(f,g) = \sum_\ell \lambda_\ell (f_\ell - g_\ell)^2 \tag{B.4.2}$$

$$\|K[f]\|_2^2 = \sum_\ell \lambda_\ell^2 f_\ell^2 \tag{B.4.3}$$

$$\sum_\ell \lambda_\ell f_\ell g_\ell = \langle fK[g]\rangle = \langle K[f]g\rangle \tag{B.4.4}$$

$$\sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} h_{\ell\ell'} f_\ell g_{\ell'} = \langle hK[f]K[g]\rangle \tag{B.4.5}$$

$$\sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} g_{\ell\ell'} f_{\ell\ell'} = \sum_\ell \lambda_\ell \langle f e_\ell K[g e_\ell]\rangle. \tag{B.4.6}$$

Suppose that f, g are probability densities with respect to μ that are bounded by C. Then

$$0 \le \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} g_{\ell\ell'} f_{\ell\ell'} \le C^2 \|\lambda\|_2^2. \tag{B.4.7}$$

Proof. We prove each claim, starting with (B.4.1). Clearly

$$\begin{aligned}
\|\theta_f\|^2_{\mathcal{H}_K} &= \|K[f]\|^2_{\mathcal{H}_K} \\
&= \left\| \int K(x,\cdot) f(x) \mathrm{d}\mu(x) \right\|^2_{\mathcal{H}_K} \\
&= \iint \langle K(x,\cdot), K(y,\cdot)\rangle_{\mathcal{H}_K} f(x) f(y) \mathrm{d}\mu(x) \mathrm{d}\mu(y) \\
&= \iint K(x,y) f(x) f(y) \mathrm{d}\mu(x) \mathrm{d}\mu(y) \\
&= \sum_\ell \lambda_\ell f_\ell^2
\end{aligned}$$

as required. The second claim (B.4.2) follows immediately from (B.4.1) by definition. For (B.4.3) by orthogonality we have

$$\begin{aligned}
\|K[f]\|_2^2 &= \| \sum_\ell \lambda_\ell f_\ell e_\ell \|_2^2 \\
&= \sum_\ell \lambda_\ell^2 f_\ell^2.
\end{aligned}$$

For (B.4.4) by the definition of $K[\cdot]$ we have

$$\begin{aligned}
\sum_\ell \lambda_\ell f_\ell g_\ell &= \left\langle \left(\sum_\ell \lambda_\ell f_\ell e_\ell \right) g \right\rangle \\
&= \langle K[f] g \rangle.
\end{aligned}$$

For (B.4.5) we can write

$$\begin{aligned}\sum_{\ell\ell'}\lambda_\ell\lambda_{\ell'}h_{\ell\ell'}f_\ell g_{\ell'} &= \sum_\ell \lambda_\ell f_\ell \left\langle \left(\sum_{\ell'}\lambda_{\ell'}g_{\ell'}e_{\ell'}\right) he_\ell \right\rangle \\ &= \sum_\ell \lambda_\ell f_\ell \langle K[g]he_\ell\rangle \\ &= \langle K[g]hK[f]\rangle.\end{aligned}$$

Finally, for (B.4.6) we have

$$\begin{aligned}\sum_{\ell\ell'}\lambda_\ell\lambda_{\ell'}f_{\ell\ell'}g_{\ell\ell'} &= \sum_\ell \lambda_\ell \left\langle \left(\sum_{\ell'}\lambda_{\ell'}g_{\ell\ell'}e_{\ell'}\right) fe_\ell \right\rangle \\ &= \sum_\ell \lambda_\ell \langle K[ge_\ell]fe_\ell\rangle.\end{aligned}$$

Suppose now that f, g are probability densities with respect to μ that are bounded by $C > 0$. Let X, Y be independent random variables following the densities f, g. Then

$$\begin{aligned}\sum_{\ell\ell'}\lambda_\ell\lambda_{\ell'}f_{\ell\ell'}g_{\ell\ell'} &= \mathbb{E}\left[\left(\sum_\ell \lambda_\ell e_\ell(X)e_\ell(Y)\right)^2\right] \\ &\le C^2\int_{\mathcal{X}}\int_{\mathcal{X}}\left(\sum_\ell \lambda_\ell e_\ell(x)e_\ell(y)\right)^2 \mathrm{d}\mu(x)\mathrm{d}\mu(y) \\ &= C^2\|\lambda\|_2^2\end{aligned}$$

as claimed, where we used that the e_ℓ are orthonormal. □

B.4.2 Mean and Variance Computation

We take $\pi = \delta/2$. Our statistic reads

$$\begin{aligned}-T(X,Y,Z) + \gamma(X,Y,\pi) &= \langle \theta_{\widehat{P}_Z} - (\bar{\pi}\theta_{\widehat{P}_X} + \pi\theta_{\widehat{P}_Y}), \theta_{\widehat{P}_X} - \theta_{\widehat{P}_Y}\rangle_{u,\mathcal{H}_K} \\ &= \underbrace{\frac{1}{nm}\sum_{ij}k(X_i,Z_j)}_{\mathsf{I}} - \underbrace{\frac{1}{nm}\sum_{ij}k(Y_i,Z_j)}_{\mathsf{II}} - \underbrace{\frac{2\bar{\pi}}{n(n-1)}\sum_{i<i'}k(X_i,X_{i'})}_{\mathsf{III}} \\ &\quad + \underbrace{\frac{2\pi}{n(n-1)}\sum_{i<i'}k(Y_i,Y_{i'})}_{\mathsf{IV}} + \underbrace{\frac{\bar{\pi}-\pi}{n^2}\sum_{ij}k(X_i,Y_j)}_{\mathsf{V}}.\end{aligned}$$

Recall that $\nu = \arg\min_{\nu'\in\mathbb{R}}\mathrm{MMD}(P_Z, \bar{\nu}'P_X + \nu'P_Y)$. Let us write $z = \bar{\nu}x + \nu y + r$ for $1-\bar{\nu} = \nu$, where the residual term is denoted as $r \in L^2(\mu)$. Let $\theta_r = \int r(t)K(t,\cdot)\mu(\mathrm{d}t)$ be the

mean embedding of r. Under both hypotheses we assume that $\|\theta_r\|_{\mathcal{H}_K} \leq R \cdot \mathrm{MMD}(P_X, P_Y)$, moreover $\langle \theta_r, \theta_{P_Y} - \theta_{P_X} \rangle_{\mathcal{H}_K} = 0$ by the definition of ν. We look at each of the $5 + \binom{5}{2} = 15$ terms of the variance separately.

$$\mathrm{var}(\mathsf{I}) = \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Big\{ n(n-1)m(z_{\ell\ell'} - z_\ell z_{\ell'}) x_\ell x_{\ell'} + nm(m-1)(x_{\ell\ell'} - x_\ell x_{\ell'}) z_\ell z_{\ell'} + nm(x_{\ell\ell'} z_{\ell\ell'} - x_\ell x_{\ell'} z_\ell z_{\ell'}) \Big\}$$

$$\mathrm{var}(\mathsf{II}) = \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Big\{ n(n-1)m(z_{\ell\ell'} - z_\ell z_{\ell'}) y_\ell y_{\ell'} + nm(m-1)(y_{\ell\ell'} - y_\ell y_{\ell'}) z_\ell z_{\ell'} + nm(y_{\ell\ell'} z_{\ell\ell'} - y_\ell y_{\ell'} z_\ell z_{\ell'}) \Big\}$$

$$\mathrm{var}(\mathsf{III}) = \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Big\{ \binom{n}{2} (x_{\ell\ell'}^2 - x_\ell^2 x_{\ell'}^2) + (\binom{n}{2}^2 - \binom{n}{2} - \binom{4}{2}\binom{n}{4}) (x_{\ell\ell'} - x_\ell x_{\ell'}) x_\ell x_{\ell'} \Big\}$$

$$\mathrm{var}(\mathsf{IV}) = \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Big\{ \binom{n}{2} (y_{\ell\ell'}^2 - y_\ell^2 y_{\ell'}^2) + (\binom{n}{2}^2 - \binom{n}{2} - \binom{4}{2}\binom{n}{4}) (y_{\ell\ell'} - y_\ell y_{\ell'}) y_\ell y_{\ell'} \Big\}$$

$$\mathrm{var}(\mathsf{V}) = \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Big\{ n^2(n-1)(y_{\ell\ell'} - y_\ell y_{\ell'}) x_\ell x_{\ell'} + n^2(n-1)(x_{\ell\ell'} - x_\ell x_{\ell'}) y_\ell y_{\ell'} + n^2(x_{\ell\ell'} y_{\ell\ell'} - x_\ell x_{\ell'} y_\ell y_{\ell'}) \Big\}$$

For the cross terms we obtain

$$\begin{aligned}
\mathrm{Cov}(\mathsf{I},\mathsf{II}) &= \sum_{\ell\ell'} \lambda_\ell\lambda_{\ell'} n^2 m(z_{\ell\ell'} - z_\ell z_{\ell'}) x_\ell y_{\ell'} \\
\mathrm{Cov}(\mathsf{I},\mathsf{III}) &= \sum_{\ell\ell'} \lambda_\ell\lambda_{\ell'} n(n-1) m(x_{\ell\ell'} - x_\ell x_{\ell'}) z_\ell x_{\ell'} \\
\mathrm{Cov}(\mathsf{I},\mathsf{IV}) &= 0 \\
\mathrm{Cov}(\mathsf{I},\mathsf{V}) &= \sum_{\ell\ell'} \lambda_\ell\lambda_{\ell'} n^2 m(x_{\ell\ell'} - x_\ell x_{\ell'}) z_\ell y_{\ell'} \\
\mathrm{Cov}(\mathsf{II},\mathsf{III}) &= 0 \\
\mathrm{Cov}(\mathsf{II},\mathsf{IV}) &= \sum_{\ell\ell'} \lambda_\ell\lambda_{\ell'} n(n-1) m(y_{\ell\ell'} - y_\ell y_{\ell'}) z_\ell y_{\ell'} \\
\mathrm{Cov}(\mathsf{II},\mathsf{V}) &= \sum_{\ell\ell'} \lambda_\ell\lambda_{\ell'} n^2 m(y_{\ell\ell'} - y_\ell y_{\ell'}) z_\ell x_{\ell'} \\
\mathrm{Cov}(\mathsf{III},\mathsf{IV}) &= 0 \\
\mathrm{Cov}(\mathsf{III},\mathsf{V}) &= \sum_{\ell\ell'} \lambda_\ell\lambda_{\ell'} n^2 (n-1) (x_{\ell\ell'} - x_\ell x_{\ell'}) x_\ell y_{\ell'} \\
\mathrm{Cov}(\mathsf{IV},\mathsf{V}) &= \sum_{\ell\ell'} \lambda_\ell\lambda_{\ell'} n^2 (n-1) (y_{\ell\ell'} - y_\ell y_{\ell'}) y_\ell x_{\ell'}.
\end{aligned}$$

Note that $\binom{n}{2}^2 - \binom{n}{2} - \binom{n}{2}\binom{n}{4} = n(n-1)^2 - n(n-1)$. Collecting terms, and simplifying, we get the coefficient of the $\frac{1}{n}$ term:

$$\begin{aligned}
\mathrm{Coef}\left(\frac{1}{n}\right) = \sum_{\ell,\ell'} \lambda_\ell\lambda_{\ell'} \Bigg(& \underbrace{(x_{\ell\ell'} - x_\ell x_{\ell'}) z_\ell z_{\ell'}}_{\mathrm{var}(\mathsf{I})} + \underbrace{(y_{\ell\ell'} - y_\ell y_{\ell'}) z_\ell z_{\ell'}}_{\mathrm{var}(\mathsf{II})} + \underbrace{4\bar{\pi}^2 (x_{\ell\ell'} - x_\ell x_{\ell'}) x_\ell x_{\ell'}}_{\mathrm{var}(\mathsf{III})} \\
& + \underbrace{4\pi^2 (y_{\ell\ell'} - y_\ell y_{\ell'}) y_\ell y_{\ell'}}_{\mathrm{var}(\mathsf{IV})} + \underbrace{(\bar{\pi}-\pi)^2 (y_{\ell\ell'} - y_\ell y_{\ell'}) x_\ell x_{\ell'} + (\bar{\pi}-\pi)^2 (x_{\ell\ell'} - x_\ell x_{\ell'}) y_\ell y_{\ell'}}_{\mathrm{var}(\mathsf{V})} \\
& - \underbrace{4\bar{\pi} (x_{\ell\ell'} - x_\ell x_{\ell'}) z_\ell x_{\ell'}}_{\mathrm{Cov}(\mathsf{I},\mathsf{III})} + \underbrace{2(\bar{\pi}-\pi)(x_{\ell\ell'} - x_\ell x_{\ell'}) z_\ell y_{\ell'}}_{\mathrm{Cov}(\mathsf{I},\mathsf{V})} \\
& - \underbrace{4\pi (y_{\ell\ell'} - y_\ell y_{\ell'}) z_\ell y_{\ell'}}_{\mathrm{Cov}(\mathsf{II},\mathsf{IV})} - \underbrace{2(\bar{\pi}-\pi)(y_{\ell\ell'} - y_\ell y_{\ell'}) z_\ell x_{\ell'}}_{\mathrm{Cov}(\mathsf{II},\mathsf{V})} \\
& - \underbrace{4\bar{\pi}(\bar{\pi}-\pi)(x_{\ell\ell'} - x_\ell x_{\ell'}) x_\ell y_{\ell'}}_{\mathrm{Cov}(\mathsf{III},\mathsf{V})} + \underbrace{4\pi(\bar{\pi}-\pi)(y_{\ell\ell'} - y_\ell y_{\ell'}) y_\ell x_{\ell'}}_{\mathrm{Cov}(\mathsf{IV},\mathsf{V})} \Bigg).
\end{aligned}$$

After expanding z_ℓ as $z_\ell = \bar{\nu} x_\ell + \nu y_\ell + r_\ell$, we split the calculation into multiple parts to simplify it. First, we focus on terms that are multiplied by $(x_{\ell\ell'} - x_\ell x_{\ell'})$ and do not contain r_ℓ or $r_{\ell'}$. Using Lemma B.4.1 extensively and the fact that $\bar{\pi} = 1 - \pi, \bar{\nu} = 1 - \nu$, we find that

the sum of these terms equals

$$
\begin{aligned}
&\bar{\nu}^2\langle xK[x]^2\rangle + \nu^2\langle xK[y]^2\rangle + 2\bar{\nu}\nu\langle xK[x]K[y]\rangle - \bar{\nu}^2\langle xK[x]\rangle^2 - \nu^2\langle xK[y]\rangle^2 - 2\bar{\nu}\nu\langle xK[x]\rangle\langle xK[y]\rangle \\
&+ 4\bar{\pi}^2\langle xK[x]^2\rangle - 4\bar{\pi}^2\langle xK[x]\rangle^2 + (\bar{\pi}-\pi)^2\langle xK[y]^2\rangle - (\bar{\pi}-\pi)^2\langle xK[y]\rangle^2 \\
&- 4\bar{\pi}\bar{\nu}\langle xK[x]^2\rangle - 4\bar{\pi}\nu\langle xK[x]K[y]\rangle + 4\bar{\pi}\bar{\nu}\langle xK[x]\rangle^2 + 4\bar{\pi}\nu\langle xK[x]\rangle\langle xK[y]\rangle \\
&+ 2(\bar{\pi}-\pi)\bar{\nu}\langle xK[x]K[y]\rangle + 2(\bar{\pi}-\pi)\nu\langle xK[y]^2\rangle - 2(\bar{\pi}-\pi)\bar{\nu}\langle xK[x]\rangle\langle xK[y]\rangle \\
&- 2(\bar{\pi}-\pi)\nu\langle xK[y]\rangle^2 - 4\bar{\pi}(\bar{\pi}-\pi)\langle xK[x]K[y]\rangle + 4\bar{\pi}(\bar{\pi}-\pi)\langle xK[x]\rangle\langle xK[y]\rangle \\
=&(\bar{\nu}-2\bar{\pi})^2\Big(\langle xK[x-y]^2\rangle - \langle xK[x-y]\rangle^2\Big) \\
\le& C\,\|\lambda\|_\infty \,\mathrm{MMD}^2(P_X, P_Y).
\end{aligned}
$$

Similarly, the terms involving $(y_{\ell\ell'} - y_\ell y_{\ell'})$ but not r_ℓ or $r_{\ell'}$ sum up to the quantity

$$
(\nu-2\pi)^2\Big(\langle yK[x-y]^2\rangle - \langle yK[x-y]\rangle^2\Big) \le C\|\lambda\|_\infty \,\mathrm{MMD}^2(P_X, P_Y).
$$

Next, collecting the terms involving both $(x_{\ell\ell'} - x_\ell x_{\ell'})$ and r_ℓ or $r_{\ell'}$ we get

$$
\begin{aligned}
&2\bar{\nu}\langle xK[r]K[x]\rangle + 2\nu\langle xK[r]K[y]\rangle + \langle xK[r]^2\rangle - 2\bar{\nu}\langle xK[x]\rangle\langle xK[r]\rangle - 2\nu\langle xK[y]\rangle\langle xK[r]\rangle \\
&- \langle xK[r]\rangle^2 - 4\bar{\pi}\langle xK[x]K[r]\rangle + 4\bar{\pi}\langle xK[x]\rangle\langle xK[r]\rangle \\
&+ 2(\bar{\pi}-\pi)\langle xK[y]K[r]\rangle - 2(\bar{\pi}-\pi)\langle xK[y]\rangle\langle xK[r]\rangle \\
=&\,2(\bar{\nu}-2\bar{\pi})\Big(\langle xK[r]K[x-y]\rangle - \langle xK[r]\rangle\langle xK[x-y]\rangle\Big) + \langle xK[r]^2\rangle - \langle xK[r]\rangle^2 \\
\lesssim& C\,\|\lambda\|_\infty(R+R^2)\,\mathrm{MMD}^2(P_X, P_Y).
\end{aligned}
$$

Finally, collecting the terms involving both $(y_{\ell\ell'} - y_\ell y_{\ell'})$ and r_ℓ or $r_{\ell'}$ we get

$$
\begin{aligned}
&2(\nu-2\pi)\Big(\langle yK[r]K[y-x]\rangle - \langle yK[r]\rangle\langle yK[y-x]\rangle\Big) + \langle yK[r]^2\rangle - \langle yK[r]\rangle^2 \\
\lesssim& C\|\lambda\|_\infty(R+R^2)\,\mathrm{MMD}^2(P_X, P_Y).
\end{aligned}
$$

Similarly we get

$$
\begin{aligned}
\mathrm{Coef}\left(\frac{1}{m}\right) &= \sum_{\ell\ell'}\lambda_\ell\lambda_{\ell'}\left(\underbrace{(z_{\ell\ell'} - z_\ell z_{\ell'})x_\ell x_{\ell'}}_{\mathrm{var(I)}} + \underbrace{(z_{\ell\ell'} - z_\ell z_{\ell'})y_\ell y_{\ell'}}_{\mathrm{var(I)}} + \underbrace{2(z_{\ell\ell'} - z_\ell z_{\ell'})x_\ell y_{\ell'}}_{\mathrm{Cov(I,II)}}\right) \\
&= \langle zK[x-y]^2\rangle - \langle zK[x-y]\rangle^2 \\
&\lesssim C\|\lambda\|_\infty \,\mathrm{MMD}^2(P_X, P_Y).
\end{aligned}
$$

The remaining coefficients don't rely on subtle cancellations, and simple bounds yield

$$\begin{aligned}
\mathrm{Coef}\left(\frac{1}{n(n-1)}\right) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Bigg(\underbrace{4\bar{\pi}^2 \left(\frac{1}{2}(x_{\ell\ell'}^2 - x_\ell^2 x_{\ell'}^2) - (x_{\ell\ell'} - x_\ell x_{\ell'}) x_\ell x_{\ell'} \right)}_{\mathrm{var}(\mathsf{III})} \\
&\qquad + \underbrace{4\pi^2 \left(\frac{1}{2}(y_{\ell\ell'}^2 - y_\ell^2 y_{\ell'}^2) - (y_{\ell\ell'} - y_\ell y_{\ell'}) y_\ell y_{\ell'} \right)}_{\mathrm{var}(\mathsf{IV})} \Bigg) \\
&\lesssim C^2 \|\lambda\|_2^2 \\
\mathrm{Coef}\left(\frac{1}{nm}\right) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Bigg(\underbrace{-(z_{\ell\ell'} - z_\ell z_{\ell'}) x_\ell x_{\ell'} - (x_{\ell\ell'} - x_\ell x_{\ell'}) z_\ell z_{\ell'} + (x_{\ell\ell'} z_{\ell\ell'} - x_\ell x_{\ell'} z_\ell z_{\ell'})}_{\mathrm{var}(\mathsf{I})} \\
&\qquad - \underbrace{(z_{\ell\ell'} - z_\ell z_{\ell'}) y_\ell y_{\ell'} - (y_{\ell\ell'} - y_\ell y_{\ell'}) z_\ell z_{\ell'} + (y_{\ell\ell'} z_{\ell\ell'} - y_\ell y_{\ell'} z_\ell z_{\ell'})}_{\mathrm{var}(\mathsf{I})} \Bigg) \\
&\lesssim C^2 \|\lambda\|_2^2 \\
\mathrm{Coef}\left(\frac{1}{n^2}\right) &= \\
&\sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Bigg(\underbrace{(\bar{\pi} - \pi) \Bigg(- (y_{\ell\ell'} - y_\ell y_{\ell'}) x_\ell x_{\ell'} - (x_{\ell\ell'} - x_\ell x_{\ell'}) y_\ell y_{\ell'} + (x_{\ell\ell'} y_{\ell\ell'} - x_\ell x_{\ell'} y_\ell y_{\ell'}) \Bigg)}_{\mathrm{var}(\mathsf{V})} \Bigg) \\
&\lesssim C^2 \|\lambda\|_2^2.
\end{aligned}$$

Summarizing, we've found that

$$\begin{aligned}
\mathrm{var}(T(X,Y,Z) - \gamma(X,Y,\pi)) \lesssim &\left(\frac{1}{n} + \frac{1}{m}\right) C \|\lambda\|_\infty (1 + R^2) \,\mathrm{MMD}^2(P_X, P_Y) \\
&+ \left(\frac{1}{n^2} + \frac{1}{nm}\right) C^2 \|\lambda\|_2^2.
\end{aligned} \tag{B.4.8}$$

Using that $\langle \theta_r, \theta_{P_Y} - \theta_{P_X} \rangle_{\mathcal{H}_K} = 0$, we compute the expectation to be

$$\mathbb{E}\left[-T(X,Y,Z) + \gamma(X,Y,\pi)\right] = (\pi - \nu)\,\mathrm{MMD}^2(P_X, P_Y).$$

Taking $\pi := \delta/2$ and applying Chebyshev's inequality shows that there exists a universal constant $c > 0$, such that the testing problem is possible at constant error probability (say $\alpha = 5\%$), provided that the sample sizes m, n satisfy the following inequalities:

$$\begin{aligned}
\min\{m, n\} &\geq c \frac{C\|\lambda\|_\infty (1 + R^2)}{\delta^2 \epsilon^2} \\
\min\{n, \sqrt{nm}\} &\geq c \frac{C\|\lambda\|_2}{\delta \epsilon^2}.
\end{aligned}$$

By repeated sample splitting and majority voting (see Appendix B.3), we can boost the success probability of this test to the desired level $1-\alpha$ by incurring a multiplicative $\Theta(\log(1/\alpha))$ factor on the sample sizes n, m, which yields the desired result.

B.5 Proof of Theorem 3.3.2

B.5.1 Information theoretic tools

Our lower bounds rely on the method of two fuzzy hypotheses [198]. Given a measurable space $\mathcal{S}$, let $\mathcal{M}(\mathcal{S})$ denote the set of all probability measures on $\mathcal{S}$. We call subsets $H \subseteq \mathcal{M}(\mathcal{S})$ hypotheses. The following is the main technical result that our proofs rely on.

Lemma B.5.1. *Take hypotheses $H_0, H_1 \subseteq \mathcal{M}(\mathcal{S})$ and $P_0, P_1 \in \mathcal{M}(\mathcal{S})$ random with $\mathbb{P}(P_i \in H_i) = 1$. Then*

$$\inf_{\psi} \max_{i=0,1} \sup_{P \in H_i} P(\psi \neq i) \geq \frac{1}{2}\left(1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1)\right),$$

where the infimum is over all tests $\psi : \mathcal{X} \to \{0,1\}$.

Proof. For any ψ

$$\begin{aligned}\max_{i=0,1} \sup_{\mathbb{P}_i \in H_i} \mathbb{P}_i(\psi \neq i) &\geq \frac{1}{2} \sup_{\mathbb{P}_i \in H_i} \left(\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0)\right) \\ &\geq \frac{1}{2}\mathbb{E}\Big[P_0(\psi = 1) + P_1(\psi = 0)\Big].\end{aligned}$$

Optimizing over ψ we get that the RHS above is equal to $\frac{1}{2}(1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1))$ as required. □

Therefore, to prove a lower bound on the minimax sample complexity of testing with total error probability α, we just need to construct two random measures $P_i \in H_i$ such that $1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) = \Omega(\alpha)$. In our proofs we also use the following standard results on f-divergences.

Lemma B.5.2 ([168, Section 7])**.** *For any probability distributions P, Q the inequalities*

$$1 - \mathsf{TV}(P, Q) \geq \frac{1}{2}\exp(-\mathsf{KL}(P\|Q)) \geq \frac{1}{2}\frac{1}{1 + \chi^2(P\|Q)}$$

hold.

Lemma B.5.3 (Chain rule for χ^2-divergence)**.** *Let $P_{X,Y}, Q_{X,Y}$ be probability measures such that the marginals on X are equal ($P_X = Q_X$). Then*

$$\chi^2(P_{X,Y}\|Q_{X,Y}) = \chi^2(P_{Y|X}\|Q_{Y|X}|P_X).$$

Proof. Let $P_{X,Y}, Q_{X,Y}$ have densities p, q with respect to some μ. Then, by some abuse of notation, we have

$$\begin{aligned}\chi^2(P_{X,Y}\|Q_{X,Y}) &= -1 + \int \frac{p(x,y)^2}{q(x,y)}\mathrm{d}\mu(x,y) \\ &= -1 + \int \frac{p(y|x)^2 p(x)}{q(y|x)}\mathrm{d}\mu(x,y) \\ &= \int p(x) \int \left(\frac{p(y|x)^2}{q(y|x)} - 1\right)\mathrm{d}\mu(y,x) \\ &= \chi^2(P_{Y|X}\|Q_{Y|X}|P_X).\end{aligned}$$

□

B.5.2 Constructing hard instances

Recall that in the statement of Theorem 3.3.2, we assume that $\mu(\mathcal{X}) = 1$, $\sup_{x\in\mathcal{X}} K(x,x) \le 1$ and $\int K(x,y)\mu(\mathrm{d}x) \equiv \lambda_1$. Let $f_0 \equiv 1$ and for each $\eta \in \{\pm 1\}^{\mathbb{N}}$ define

$$f_\eta = 1 + \epsilon \underbrace{\sum_{j\ge 2} \rho_j \eta_j e_j}_{=:g_\eta} \tag{B.5.1}$$

where $\{\rho_j\}_{j\ge 2}$ is chosen as $\rho_j = \mathbb{1}\{2 \le j \le J\}\sqrt{\lambda_j}/\|\lambda\|_{2,J}$, where we define $\|\lambda\|_{2,J} = \sqrt{\sum_{2\le j\le J}\lambda_j^2}$ for some $J \ge 2$. Notice that $\int f_\eta(x)\mu(\mathrm{d}x) = \mu(\mathcal{X}) = 1$ due to orthogonality of the eigenfunctions. Assume from here on that J is chosen so that for all η we have $f_\eta(x) \ge 1/2$ for all $x \in \mathcal{X}$. This makes f_η into a valid probability density with respect to the base measure μ. Before continuing, we prove the following Lemma, which gives a lower bound on the maximal J for which $f_\eta \ge 1/2$ for all η.

Lemma B.5.4. *$J \le J_\epsilon^\star$ holds provided $2\epsilon\sqrt{J-1} \le \|\lambda\|_{2J}$.*

Proof of Lemma B.5.4. Notice that

$$\|e_j\|_\infty = \sup_{x\in\mathcal{X}} \langle K(x,\cdot), e_j\rangle_{\mathcal{H}} \le \sup_{x\in\mathcal{X}} \|K(x,\cdot)\|_{\mathcal{H}}\|e_j\|_{\mathcal{H}} \le \frac{1}{\sqrt{\lambda_j}}, \tag{B.5.2}$$

where we use $\|K(x,\cdot)\|_{\mathcal{H}} = \sqrt{K(x,x)}$. We have

$$\begin{aligned}\|g_\eta\|_\infty &= \epsilon\|\sum_{j\ge 2}\rho_j\eta_j e_j\|_\infty = \epsilon \sup_{x\in\mathcal{X}}\langle K(x,\cdot), \sum_{j\ge 2}\rho_j\eta_j e_j\rangle_{\mathcal{H}} \\ &\le \epsilon\|\sum_{j\ge 2}\rho_j\eta_j e_j\|_{\mathcal{H}} = \epsilon\sqrt{\sum_{j\ge 2}\rho_j^2/\lambda_j} = \frac{\epsilon\sqrt{J-1}}{\|\lambda\|_{2,J}},\end{aligned}$$

and the result follows. □

Note that Lemma B.5.4 immediately gives us a proof of Corollary 3.3.3.

Proof of Corollary 3.3.3. Suppose that J is such that $\sum_{j=2}^{J} \lambda_j^2 \geq c^2 \|\lambda\|_2^2$. Then, by Lemma B.5.4, if $\epsilon \leq \|\lambda\|_{2J}/(2\sqrt{J-1})$ then $J \leq J_\epsilon^\star$. By assumption, this is implied by the inequality $\epsilon \leq c\|\lambda\|_2/(2\sqrt{J-1})$, and the result follows. □

Continuing with our proof, note that by construction we have

$$\mathrm{MMD}^2(f_0, f_\eta) = \sum_{j \geq 2} \lambda_j \rho_j^2 = \epsilon^2, \quad \forall \eta \in \{\pm 1\}^{\mathbb{N}}. \tag{B.5.3}$$

Lower Bound on m

Again, we apply Lemma B.5.1 with the new (deterministic) construction

$$P_0 = f_0^{\otimes n} \otimes (1 + \epsilon e_2/\sqrt{\lambda_2})^{n} \otimes (1 + \delta\epsilon e_2/\sqrt{\lambda_2})^{\otimes m}, \qquad P_1 = f_0^{\otimes n} \otimes (1 + \epsilon e_2/\sqrt{\lambda_2})^{n} \otimes f_0^{\otimes m}, \tag{B.5.4}$$

where we write $f_{\mathbb{1}} = f_{(1,1,\dots)}$ and similarly for $g_{\mathbb{1}}$. By the data-processing inequality for χ^2-divergence (also by Lemma B.5.3), we may drop the first $2n$ coordinates and obtain

$$\begin{aligned}
\chi^2(\mathbb{E}P_0, \mathbb{E}P_1) &= \chi^2((1 + \delta\epsilon e_2/\sqrt{\lambda_2})^{\otimes m} \| f_0^{\otimes m}) \\
&= \left(1 + \delta^2\epsilon^2/\lambda_2\right)^m - 1 \\
&\leq \exp\left(\frac{\delta^2\epsilon^2 m}{\lambda_2}\right) - 1.
\end{aligned}$$

By Lemma B.5.2 we

$$1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) \gtrsim \frac{1}{\chi^2(\mathbb{E}P_0, \mathbb{E}P_1) - 1} \geq \exp(-\delta^2\epsilon^2 m) \overset{!}{=} \Omega(\alpha).$$

The lower bound $m \gtrsim \lambda_2 \log(1/\alpha)/(\delta\epsilon)^2$ now follows readily.

Lower Bound on n

Once again, we apply Lemma B.5.1 to the new construction

$$P_0 = f_0^{\otimes n} \otimes f_\eta^{\otimes n} \otimes f_0^{\otimes m}, \qquad P_1 = f_\eta^{\otimes n} \otimes f_0^{\otimes n} \otimes f_0^{\otimes m}, \tag{B.5.5}$$

where we put a uniform prior on $\eta \in \{\pm 1\}^{\mathbb{N}}$ as before. Using the subadditivity of total variation under products, we compute

$$\begin{aligned}
\mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) &= \mathsf{TV}(f_0^{\otimes n} \otimes \mathbb{E}f_\eta^{\otimes n}, \mathbb{E}[f_\eta^{\otimes n}] \otimes f_0^{\otimes n}) \\
&\leq 2\mathsf{TV}(\mathbb{E}f_\eta^{\otimes n}, f_0^{\otimes n}).
\end{aligned}$$

Just as in Appendix B.5.2 we upper bound by the χ^2-divergence to get

$$\begin{aligned}\chi^2(\mathbb{E}f_\eta^{\otimes n}\|f_0^{\otimes n}) &= -1 + \mathbb{E}_{\eta\eta'}\int\prod_{i=1}^n(f_\eta(x_i)f_{\eta'}(x_i))\mu(\mathrm{d}x_1)\dots\mu(\mathrm{d}x_n)\\ &\le -1 + \mathbb{E}\exp(n\epsilon^2\sum_{j\ge2}\rho_j^2\eta_j\eta_j')\\ &= -1 + \prod_{j\ge2}\cosh(n\epsilon^2\rho_j^2)\\ &\le -1 + \exp(n^2\epsilon^4\sum_{j\ge2}\rho_j^4)\\ &= -1 + \exp(n^2\epsilon^4/\|\lambda\|_{2,J}^2).\end{aligned}$$

Again, by Lemma B.5.2 we obtain

$$1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) \gtrsim \frac{1}{\chi^2(\mathbb{E}P_0\|\mathbb{E}P_1) - 1} \ge \exp(-n^2\epsilon^4/\|\lambda\|_{2,J}^2) \overset{!}{=} \Omega(\alpha).$$

The lower bound $n \gtrsim \sqrt{\log(1/\alpha)}\|\lambda\|_{2,J}/\epsilon^2$ now follows readily.

Lower Bound on $m \cdot n$

We take a uniform prior on η and consider the random measures

$$P_0 = f_0^{\otimes n} \otimes f_\eta^{\otimes n} \otimes ((1-\delta)f_0 + \delta f_\eta)^{\otimes m} \qquad \text{and} \qquad P_1 = f_0^{\otimes n} \otimes f_\eta^{\otimes n} \otimes f_0^{\otimes m}. \tag{B.5.6}$$

Our goal is to apply Lemma B.5.1 to P_0, P_1. Notice that $(1-\delta)f_0 + \delta f_\eta = 1 + \delta\epsilon g_\eta$. Let us write X, Y, Z for the marginals first n, second n and last m coordinates of P_0 and P_1. By the data processing inequality and the chain rule Lemma B.5.3 we have

$$\begin{aligned}\chi^2(\mathbb{E}P_0\|\mathbb{E}P_1) &= \chi^2((\mathbb{E}P_0)_{Y,Z}\|(\mathbb{E}P_1)_{Y,Z})\\ &= \chi^2((\mathbb{E}P_0)_{Z|Y}\|(\mathbb{E}P_1)_{Z|Y}|(\mathbb{E}P_0)_Y)\\ &= \mathbb{E}\chi^2\left(\mathbb{E}\left[(1+\delta\epsilon g_\eta)^{\otimes m}\middle| Y\right]\|f_0^{\otimes m}\right) =: (\dagger).\end{aligned}$$

Notice that the expectation inside the χ^2-divergence is with respect to η given the variables Y, or in other words, over the posterior of η with uniform prior given n observations from the density $1 + \epsilon g_\eta = f_\eta$. The outer expectation is over Y. Given Y, let η and η' be i.i.d. from said posterior. We get the bound

$$\begin{aligned}(\dagger) + 1 &\le \mathbb{E}\int\prod_{i=1}^m(1+\delta\epsilon g_\eta(x_i))(1+\delta\epsilon g_{\eta'}(x_i))\mu(\mathrm{d}x_i)\\ &= \mathbb{E}(1+\delta^2\epsilon^2\sum_{j\ge2}\rho_j^2\eta_j\eta_j')^m\\ &\le \mathbb{E}\exp(\delta^2\epsilon^2 m\sum_{j\ge2}\rho_j^2\eta_j\eta_j').\end{aligned}$$

Define the collections of variables $\eta_{-j} = \{\eta_j\}_{j\geq 2} \setminus \{\eta_j\}$ and η'_{-j} similarly. We shall prove the following claim:

$$\mathbb{E}\left[\exp(\delta^2\epsilon^2 m\rho_j^2\eta_j\eta'_j)\middle|\, \eta_{-j}\eta'_{-j}\right] \leq \exp(c\delta^2\epsilon^4(\delta^2 m^2 + mn)\rho_j^4) \tag{B.5.7}$$

for some universal constant $c > 0$. Assuming that (B.5.7) holds, by induction we can show that

$$\begin{aligned}(\dagger) + 1 &\leq \exp(c\delta^2(\delta^2 m^2 + mn)\epsilon^4 \sum_{j\geq 2} \rho_j^4) \\ &= \exp(c\delta^2(\delta^2 m^2 + mn)\epsilon^4/\|\lambda\|_{2,J}^2).\end{aligned}$$

Thus, if $mn + \delta^2 m^2 = o\left(\|\lambda\|_{2,J}^2/(\delta^2\epsilon^4)\right)$ then testing is impossible.

We now prove (B.5.7). Since the variable $\eta'_j\eta'_j$ is either 1 or -1, we have

$$\mathbb{E}\left[\exp(\delta^2\epsilon^2 m\rho_j^2\eta_j\eta'_j)\middle|\, \eta_{-j}\eta'_{-j}\right] = (e^{\delta^2\epsilon^2 m\rho_j^2} - e^{-\delta^2\epsilon^2 m\rho_j^2}) \cdot \mathbb{P}(\eta_j\eta'_j = 1|\eta_{-j}\eta'_{-j}) + e^{-\delta^2\epsilon^2 m\rho_j^2}.$$

Let us write $\eta_{\pm 1,j}$ for the vector of signs equal to η but whose j'th coordinate is ± 1 respectively. Looking at the probability above, and using the independence of η, η' given Y, we have

$$\begin{aligned}\mathbb{P}(\eta_j\eta'_j = 1|Y, \eta_{-j}, \eta'_{-j}) &= \mathbb{P}(\eta_j = 1|Y, \eta_{-j})^2 + \mathbb{P}(\eta_j = -1|Y, \eta_{-j})^2 \\ &= \frac{1}{4}\frac{(f_{\eta_{1j}}^{\otimes n}(Y))^2 + (f_{\eta_{-1j}}^{\otimes n}(Y))^2}{\left(\frac{1}{2}f_{\eta_{1j}}^{\otimes n}(Y) + \frac{1}{2}f_{\eta_{-1j}}^{\otimes n}(Y)\right)^2}.\end{aligned}$$

Taking the expectation $\mathbb{E}[\cdot\,|\eta_{-j}, \eta'_{-j}]$ and using the HM-AM inequality $(\frac{1}{2}(x+y))^{-1} \leq \frac{1}{2}(\frac{1}{x} + \frac{1}{y})$ valid for all $x, y > 0$ gives

$$\begin{aligned}\mathbb{P}(\eta_j\eta'_j = 1|\eta_{-j}, \eta'_{-j}) &= \frac{1}{4}\int \frac{(\prod_{i=1}^n f_{\eta_{1j}}(x_i))^2 + (\prod_{i=1}^n f_{\eta_{-1j}}(x_i))^2}{\frac{1}{2}\prod_{i=1}^n f_{\eta_{1j}}(x_i) + \frac{1}{2}\prod_{i=1}^n f_{\eta_{-1j}}(x_i)}\mu(\mathrm{d}x_1)\ldots\mu(\mathrm{d}x_n) \\ &\leq \frac{1}{4} + \frac{1}{8}\int\left(\frac{(\prod_{i=1}^n f_{\eta_{1j}}(x_i))^2}{\prod_{i=1}^n f_{\eta_{-1j}}(x_i)} + \frac{(\prod_{i=1}^n f_{\eta_{-1j}}(x_i))^2}{\prod_{i=1}^n f_{\eta_{1j}}(x_i)}\right)\mu(\mathrm{d}x_1)\ldots\mu(\mathrm{d}x_n) = (\star).\end{aligned}$$

Note that $f_{\eta_{1j}} = f_{\eta_{-1j}} + 2\epsilon\rho_j e_j$. Using the lower bound $f_{\eta_{\pm 1j}}(x) \geq \frac{1}{2}$ for all $x \in \mathcal{X}$ and the inequality $1 + x \leq \exp(x)$, we get

$$\begin{aligned}(\star) &\leq \frac{1}{4} + \frac{1}{8}\left[\left(1 + \int\frac{4\epsilon^2\rho_j^2 e_j^2(x)}{f_{\eta_{-1j}}(x)}\mu(\mathrm{d}x)\right)^n + \left(1 + \int\frac{4\epsilon^2\rho_j^2 e_j^2(x)}{f_{\eta_{1j}}(x)}\mu(\mathrm{d}x)\right)^n\right] \\ &\leq \frac{1}{4}(1 + e^{8\epsilon^2 n\rho_j^2}).\end{aligned}$$

Recall that $(\star)$ is a probability so $(\star) \leq 1$, and we obtain

$$(\star) \leq \frac{1}{4}(1 + e^{8\epsilon^2 n\rho_j^2 \wedge \ln 3}).$$

Putting it together and applying Lemma B.5.5 we get

$$\begin{aligned}\text{LHS of (B.5.7)} &\le (e^{\delta^2\epsilon^2 m\rho_j^2} - e^{-\delta^2\epsilon^2 m\rho_j^2})\frac{1}{4}(1 + e^{8\epsilon^2 n\rho_j^2 \wedge \ln 3}) + e^{-\delta^2\epsilon^2 m\rho_j^2} \\ &\le e^{c\delta^2\epsilon^4\rho_j^4(\delta^2 m^2 + mn)}\end{aligned}$$

for universal $c = 16 > 0$. Thus, by Lemma B.5.2 we obtain

$$1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) \gtrsim \frac{1}{\chi^2(\mathbb{E}P_0, \mathbb{E}P_1) + 1} \ge \exp(-c\delta^2\epsilon^4(\delta^2 m^2 + mn)/\|\lambda\|_{2,J}^2) \overset{!}{=} \Omega(\alpha).$$

The necessity of

$$mn + \delta^2 m^2 \gtrsim \frac{\log(1/\alpha)\|\lambda\|_{2,J}^2}{\delta^2\epsilon^4}$$

follows immediately.[2]

Lemma B.5.5. *For $a, b \ge 0$, the following inequality holds:*

$$\frac{1}{4}(e^a - e^{-a})(1 + e^{b\wedge\ln 3}) + e^{-a} \le e^{2(ab+a^2)}.$$

Proof. If $b \ge \ln 3$ or $a \ge 1$ we have:

$$\text{LHS} \le \frac{1}{4}(e^a - e^{-a})(1 + e^{\ln 3}) + e^{-a} = e^a \le e^{\frac{b}{\ln 3}a + a^2}.$$

If $b < \ln 3$ and $a < 1$, we have

$$e^b \le 1 + \frac{2}{\ln 3}b \le 1 + 2b, \quad \frac{e^a + e^{-a}}{2} \le e^{a^2}, \quad \frac{e^a - e^{-a}}{2} \le \frac{e - e^{-1}}{2}a \le 2a,$$

and then

$$\begin{aligned}\frac{1}{4}(e^a - e^{-a})(1 + e^b) + e^{-a} &= \frac{1}{2}(e^a + e^{-a}) + \frac{e^b - 1}{4}(e^a - e^{-a}) \\ &\le e^{a^2} + 2ab \\ &\le e^{a^2}(1 + 2ab) \\ &\le e^{a^2 + 2ab}\end{aligned}$$

The result follows from $\ln 3 > 1$. □

B.6 Proofs From Section 3.4

B.6.1 Computing $\hat{\sigma}$

We follow the implementation of $\widehat{\sigma}^2$ in [143]. Given $X_1, \dots, X_{n_{\mathsf{tr}}}^{\mathsf{tr}}$ sampled from P_X and $Y_1, \dots, Y_{n_{\mathsf{tr}}}^{\mathsf{tr}}$ sampled from P_Y, denote

$$H_{ij} := K(X_i^{\mathsf{tr}}, X_j^{\mathsf{tr}}) + K(Y_i^{\mathsf{tr}}, Y_j^{\mathsf{tr}}) - K(X_i^{\mathsf{tr}}, Y_j^{\mathsf{tr}}) - K(Y_i^{\mathsf{tr}}, X_j^{\mathsf{tr}}), \quad i, j \in [n_{\mathsf{tr}}]. \tag{B.6.1}$$

[2] We have $mn + m^2 \le (\sqrt{mn} + m)^2 \le 2(mn + m^2)$, so $\sqrt{mn} + m \asymp \sqrt{mn + m^2}$.

Then $\widehat{\sigma}^2$ is computed via

$$\hat{\sigma}^2(X^{n_{\text{tr}}}, Y^{n_{\text{tr}}}; K) = \frac{4}{n_{\text{tr}}^3} \sum_{i=1}^{n_{\text{tr}}} \left(\sum_{j=1}^{n_{\text{tr}}} H_{ij} \right)^2 - \frac{4}{n_{\text{tr}}^4} \left(\sum_{i=1}^{n_{\text{tr}}} \sum_{j=1}^{n_{\text{tr}}} H_{ij} \right)^2. \tag{B.6.2}$$

Note that $\hat{\sigma}^2$ being non-negative follows from the AM-GM inequality.

B.6.2 Heuristic Justification of the Objective (3.4.1)

As usual, let X, Y, Z denotes samples of sizes n, n, m from P_X, P_Y, P_Z respectively. Let us give a heuristic justification for using the training objective defined in (3.4.1) for the purpose of obtaining a kernel for LFHT/mLFHT. Note that originally it was proposed as a training objective for kernels to be used in two sample testing. Recall that our test for LFHT can be written as

$$\Psi_{1/2}(X, Y, Z) = \mathbb{1}\Big\{ T_{\mathsf{LF}} \geq 0 \Big\}$$

where

$$T_{\mathsf{LF}} = \mathrm{MMD}_u^2(\widehat{P}_Z, \widehat{P}_Y; K) - \mathrm{MMD}_u^2(\widehat{P}_Z, \widehat{P}_X; K),$$

Heuristically, to maximize the power of (mLFHT), we would like to maximize the following population quantity

$$J_{\mathsf{LF}} := \frac{\mathbb{E}_0[T_{\mathsf{LF}}] - \mathbb{E}_1[T_{\mathsf{LF}}]}{\sqrt{\mathrm{var}_0(T_{\mathsf{LF}})}}$$

where

$$\begin{aligned}
\mathbb{E}_0[T_{\mathsf{LF}}] &= \mathbb{E}_{X,Y,Z}[T_{\mathsf{LF}} | P_Z = P_X] = +\mathrm{MMD}^2(P_X, P_Y; K), \\
\mathbb{E}_1[T_{\mathsf{LF}}] &= \mathbb{E}_{X,Y,Z}[T_{\mathsf{LF}} | P_Z = P_Y] = -\mathrm{MMD}^2(P_X, P_Y; K).
\end{aligned}$$

Let $T_{\mathsf{TS}} = \mathrm{MMD}_u(\hat{P}_X, \hat{P}_Y)$ be the usual statistic that is thresholded for two-sample testing. Then, a computation analogous to that in Section B.4.2 show (cf. (B.4.8)) that

$$\begin{aligned}
\mathrm{var}_0(T_{\mathsf{LF}}) &\approx \frac{A(K, P_X, P_Y)}{n} + \frac{A(K, P_X, P_Y)}{m} + \frac{B(K, P_X, P_Y)}{n^2} + \frac{B(K, P_X, P_Y)}{mn}, \\
\mathrm{var}_0(T_{\mathsf{TS}}) &\approx \frac{A(K, P_X, P_Y)}{n} + \frac{B(K, P_X, P_Y)}{n^2}
\end{aligned}$$

for some $A(K)$ and $B(K)$. Therefore, we have approximately

$$J_{\mathsf{LF}} \approx \frac{2\,\mathrm{MMD}^2(P_X, P_Y; K)}{\sqrt{1 + \frac{n}{m}}\sqrt{\mathrm{var}_0(T_{\mathsf{TS}})}} \approx 2\sqrt{\frac{m}{m+n}}\widehat{J}(X, Y; K)$$

which only differs from our optimization objective defined in (3.4.1) by a constant factor.

Second, notice that $\frac{\mathrm{MMD}(P_X, P_Y; K)}{\sqrt{\mathrm{var}(T_{\mathsf{TS}})}}$ depends only on $P_X - P_Y$ and that $((1-\delta)P_X + \delta P_Y) - P_X \propto P_Y - P_X$, therefore it is sensible to use (3.4.1) as our training objective for is also sensible for (mLFHT), and we don't even need to observe the sample Z.

B.6.3 Proof of Proposition 3.4.1

Proof. In this proof we regard $\mathcal{D} := (X^{\mathsf{tr}}, X^{\mathsf{ev}}, Y^{\mathsf{tr}}, Y^{\mathsf{ev}})$ and the parameters of the kernel ω as fixed. Recall that we are looking at the problem mLFHT with a misspecification parameter $R = 0$ (see Theorem 3.3.1). Given a test set $\{z_i\}_{i\in[m]}$, our test statistic is $T(\{z_i\}_{i\in[m]}) = \frac{1}{m}\sum_{i=1}^{m} f(z_i)$ where

$$f(z_i) = \frac{1}{n_{\mathsf{ev}}} \sum_{j=1}^{n_{\mathsf{ev}}} \Big(K_\omega(z_i, Y_j^{\mathsf{ev}}) - K_\omega(z_i, X_j^{\mathsf{ev}}) \Big).$$

In Phase 3 of Algorithm 1, we observe the value $\widehat{T} = T(Z) = \frac{1}{m}\sum_{i=1}^{m} f(Z_i)$ and reject the null hypothesis for large values of $\widehat{T}$. Thus, the p-value is defined as

$$p = p(Z, \mathcal{D}) := \mathbb{P}_{\widetilde{Z}\sim P_X^{\otimes m}}(T(\widetilde{Z}) > \widehat{T}).$$

Phase 2 of our Algorithm 1 produces random variables $T_1, \ldots, T_k$ that all have the distribution of $T(\{\widetilde{Z}_i\}_{i\in[m]})$, so that $\mathbb{1}\{T_r \geq \widehat{T}\}$ $(r = 1, \ldots, k)$ are unbiased estimates of the p-value. However, the T_i are not independent, because they sample from the finite collection of calibration samples X^{cal}. However, as $n_{\mathsf{cal}} \to \infty$ the covariances between T_{r_1}, T_{r_2} for $r_1 \neq r_2$ tend to zero, and we obtain a consistent estimate of p. □

B.6.4 Proof of Proposition 3.4.2

Proof. The test statistic $T(X, Y, Z)$ in (3.2.3) is given by

$$T(X, Y, Z) = \frac{1}{m}\sum_{i=1}^{m} f_K(Z_i)$$

where

$$f_K(z) = \theta_{\widehat{P}_Y}(z) - \theta_{\widehat{P}_X}(z).$$

This simplifies to (consider $K(x, y) = f(x)f(y)$)

$$f_K(z) = \left(\frac{1}{n}\sum_{j=1}^{n} f(Y_j) - \frac{1}{n}\sum_{j=1}^{n} f(X_j) \right) f(z) = C(X, Y) f(z).$$

where $C(X, Y)$ does not depend on z. Therefore, for any witness function f, we obtain the desired additive test. □

B.6.5 Additive Test Statistics

In this section we prove accordingly that the test statistics of all of **MMD-M/G/O**, **SCHE**, **LBI**, **UME**, **RFM** are of the form $T_f(Z) = \frac{1}{m}\sum_{i=1}^{m} f(Z_i)$ (where f might depends on X, Y). The test is to compare $T_f(Z)$ with some threshold $\gamma(X, Y)$.

Note that in the setting of Algorithm 1, the X and Y here correspond to X^{ev} and Y^{ev}.

MMD-M/G/O As described in (3.2.3) we have

$$T_f(Z) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n}\left(K(Z_i, Y_j) - K(Z_i, X_j)\right)\right).$$

SCHE As described in Section 3.4.2 we have

$$T_f(Z) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\{\phi(Z_i) > t\}.$$

LBI As described in Section 3.4.2 we have

$$T_f(Z) = \frac{1}{m}\sum_{i=1}^{m}\log\left(\frac{\phi(Z_i)}{1-\phi(Z_i)}\right).$$

UME As described in [118], the UME statistic evaluates the squared witness function at J_q test locations $W = \{w_k\}_{k=1}^{J_q} \subset \mathcal{X}$. Formally for any two distributions P, Q we define

$$U^2(P,Q) = \|\theta_Q - \theta_P\|^2_{L^2(W)} = \frac{1}{J_q}\sum_{k=1}^{J_q}(\theta_Q(w_k) - \theta_P(w_k))^2.$$

However, we note a crucial difference that their result only considers the case of $n = m$, and their proposed estimator for $U^2(P_Z, P_X)$ can not be naturally extended to the case of $n \neq m$. Here we generalize it to $m \neq n$ where we (conveniently) use a biased estimate of their distance. Given samples X, Y, Z and a set of witness locations W, the test statistic is a (biased yet) consistent estimator of $U^2(P_Z, P_Y) - U^2(P_Z, P_X)$. Let $\psi_W(z) = \frac{1}{\sqrt{J_q}}(K(z, w_1), \ldots, K(z, w_{J_q})) \in \mathbb{R}^{|W|}$ be the "feature function," then:

$$\begin{aligned}\widehat{U}^2(Z,X) &= \left\|\frac{1}{m}\sum_{i=1}^{m}\psi_W(Z_i) - \frac{1}{n}\sum_{j=1}^{n}\psi_W(X_i)\right\|_2^2 \\ &= \left\|\frac{1}{m}\sum_{i=1}^{m}\psi_W(Z_i)\right\|_2^2 + \left\|\frac{1}{n}\sum_{j=1}^{n}\psi_W(X_i)\right\|_2^2 - \frac{2}{mn}\sum_{1\le i\le m, 1\le j\le n}\langle\psi_W(Z_i), \psi_W(X_j)\rangle\end{aligned}$$

Here $\langle\cdot,\cdot\rangle$ denotes the usual inner product. Therefore, the difference between distances is

$$\widehat{U}^2(Z,Y) - \widehat{U}^2(Z,X) = \frac{1}{m}\sum_{i=1}^{m}\left\langle\psi_W(Z_i), \frac{2}{n}\sum_{j=1}^{n}(\psi_W(X_j) - \psi_W(Y_j))\right\rangle + F(X,Y)$$

where F is sum function based only on X, Y. This is clearly an additive statistic for Z.

RFM Algorithm 1 in [170] describes a method for learning a kernel from data given a binary classification task. For convenience lets concatenate the data to $X^{\text{RFM}} = (X, Y) \in \mathbb{R}^{2n\times d}$ and labels $y^{\text{RFM}} = (\vec{0}_n, \vec{1}_n) \in \mathbb{R}^{1\times 2n}$. Given a learned kernel K, we write the Gram

matrix as $(K(X^{\mathrm{RFM}}, X^{\mathrm{RFM}}))_{i,j} = K(X_i^{\mathrm{RFM}}, X_j^{\mathrm{RFM}})$ $(1 \le i, j \le 2n)$. Let $K(X^{\mathrm{RFM}}, z)$ be a column vector with components $K(X_i^{\mathrm{RFM}}, z)$ $(1 \le i \le 2n)$. The classifier is then defined as

$$f^{\mathrm{RFM}}(z) = y^{\mathrm{RFM}} \cdot K(X^{\mathrm{RFM}}, X^{\mathrm{RFM}})^{-1} \cdot K(X^{\mathrm{RFM}}, z). \tag{B.6.3}$$

Though in [170] the kernel learned from RFM is used to construct a classifier as in Equation (B.6.3), since RFM is a feature learning method, we also apply the RFM kernel to our MMD test, namely

$$f^{\text{RFM to MMD}}(z) = \frac{1}{n} \sum_{j=1}^{n} \left(K(z, Y_j) - K(z, X_j) \right).$$

B.7 Application: Diffusion Models vs CIFAR

We defer a more fine-grained detail to our code submission, which includes executable programs (with PyTorch) once the data-generating script from DDPM has been run (see README in the ./codes/CIFAR folder).

B.7.1 Dataset Details

We use the CIFAR-10 dataset available online at https://www.cs.toronto.edu/~kriz/cifar.html, which contains 50000 colored images of size 32×32 with 10 classes. For the diffusion generated images, we use the SOTA Hugging Face model (DDPM) that can be found at https://huggingface.co/google/ddpm-CIFAR-10-32. We generated 10000 artificial images for our experiments. The code can be found at our code supplements.

For dataset balancing, we randomly shuffled the CIFAR-10 dataset and used 10000 images as data in our code. Most of our experiments are conducted with the null P_X as CIFAR images, and the alternate as $P_Y = \frac{2}{3} \cdot \mathrm{CIFAR} + \frac{1}{3} \cdot \mathrm{DDPM}$. To this end, we matched 20000 images from CIFAR to belong to the alternate hypothesis, and the remaining 30000 images to stay in the null hypothesis. For the alternate dataset, we simply sample without replacement from the $20000 + 10000$ mixture. This sampled distribution is *almost* the same as mixing (so long as the sample bank is large enough compared to the acquired data, so that each item in the alternate has close to 1/3 probability of being in DDPM, which is indeed the case).

B.7.2 Experiment Setup and Benchmarks

We use a standard deep Conv-net [144], which has been employed for SOTA GAN discriminator tasks in similar settings. It has four convolutional layers and one fully connected layer outputting the feature space of size $(300, 1)$. For SCHE and LBI, we simply added a linear layer of $(300, 2)$ after applying ReLU to the 300-dimensional layer and used the cross-entropy loss to train the network. Note that this is equivalent to first fixing the feature space and then performing logistic regression to the feature space. For kernels, we add extra trainable parameters after the 300-d feature output.

For the MMD-based tests, we simply train the kernel on the neural net and evaluate our objective. For UME, we used a slightly generalized version of the original statistic in [118]

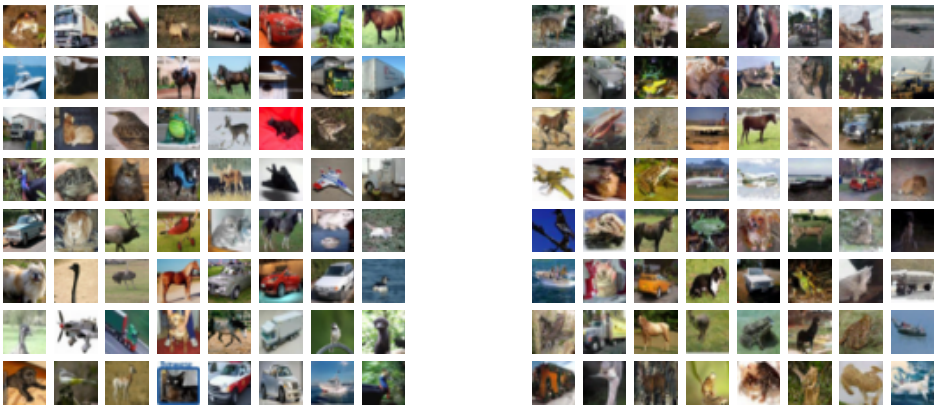


Figure B.1: Data visualization for CIFAR-10 (left) vs DDPM diffusion generated images (right)

which allows for comparison on randomly selected witness locations in the null hypothesis with $m \neq n$ (see Appendix B.6.5). The kernel is trained using our heuristic (see (3.4.1) and Appendix B.6.2), with MMD replaced by UME. The formula for UME variance can be found in [118]. For RFM, we use Algorithm 1 in [170] to learn a kernel on (stochastic batched) samples, and then use our MMD test on the trained kernel.

We use 80 training epochs for most of our code from the CNN architecture (for classifiers, this is well after interpolating the training data and roughly when validation loss stops decreasing), and a batch size of 32 which has a slight empirical benefit compared to larger batch sizes. The learning rates are tuned separately in MMD methods for optimality, whereas for classifiers they follow the discriminator's original setting from [144]. In Phase 2 of Algorithm 1, we choose $k = 1000$ for the desired precision while not compromising runtime. For each task, we run 10 independent models and report their performances as the mean and standard deviation of those 10 runs as estimates. We refer to a full set of hyper-parameters in our code implementation.

Our code is implemented in Python 3.7 (PyTorch 1.1) and was ran on an NVIDIA RTX 3080 GPU equipped with a standard torch library and dataset extensions. Our code setup for feature extraction is similar to that of [143]. For benchmark implementations, our code follows from the original code templated provided by the cited papers.

B.7.3 Sample Allocation

We make a comment on why (3.2.4) is *different* from just thresholding $\widehat{\mathrm{MMD}}^2(Z, Y^{\mathsf{tr}}) - \widehat{\mathrm{MMD}}^2(Z, X^{\mathsf{tr}})$ at 0, which was what we did in part (c) of Figure 3.3 (and hence the difference along the curve of MMD-M vs Figure 3.1). Our theory assumes that the samples are i.i.d. conditioned on the kernel being chosen already. However, in the experiments, the kernel is dependent on the training data. Therefore, to evaluate the MMD estimate (between

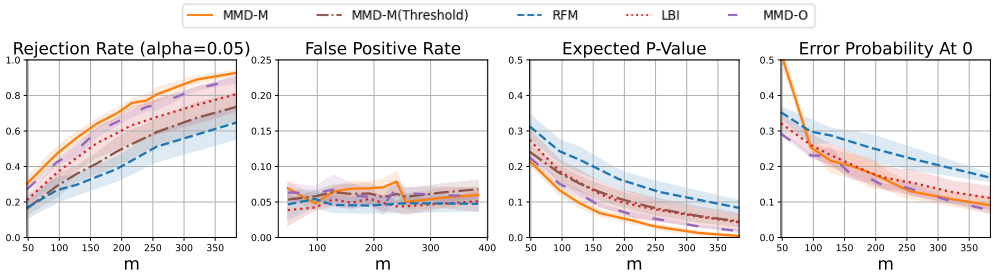


Figure B.2: Relevant plots following the setting in Figure 3.3 (in the main text) of fixing $n_{\text{tr}} = 1920$ and varying sample size m in the x-axis for the comparison with missing benchmarks. Errorbars are projected showing standard deviation across 10 runs. We replaced part (d) in Figure 3.3 (in the main text) to a sanity check in our FPR when thresholded at $\alpha = 0.05$.

experimentations), one needs extra data that does not intersect with training.

In fact, it can be experimentally shown by comparing Figure 3.1 and Figure 2(c) that doing so (while reducing the sample complexity on n_{ev}) hurts performance. Indeed, we found out that when $X^{\text{ev}}, Y^{\text{ev}}$ are non-intersecting with training, performance is (almost) always better at a cost of hurting the overall sample complexity of n.

B.7.4 Remarks on Results

Figure B.2 lists all of our benchmarks in the setting of Figure 3.3 (in the main text) on missing benchmarks, where the last figure is replaced by the false positive rate at thresholding at $\alpha = 0.05$ to verify our results. As mentioned in the main text, our MMD-M method consistently outperforms other benchmarks on both the expected p-value (of alternate) and rejection rate at $\alpha = 0.05$, while all of our tests observe a empirical false positive rate close to $\alpha = 0.05\%$ (Part (b)), showing the consistency of methods.

B.8 Application: Higgs-Boson Detection

B.8.1 Dataset Details

We use the Higgs dataset available online at http://archive.ics.uci.edu/ml/datasets/HIGGS, produced using Monte Carlo simulations [14]. The dataset is nearly balanced, containing $5,829,122$ signal instances and $5,170,877$ background instances. Each instance is a 28-dimensional vector, consisting of 28 features. The first 21 features are kinematic properties measured by the detectors in the accelerator, such as momentum and energy. The last 7 properties are *invariant masses*, derived from the first 21 features.

B.8.2 Experiment Setup and Training Models

The modified Algorithm 1 is shown in Algorithm 2 and Algorithm 3. Compared with Algorithm 2, we implement the thresholding trick (Section 3.4.3) in Algorithm 3.

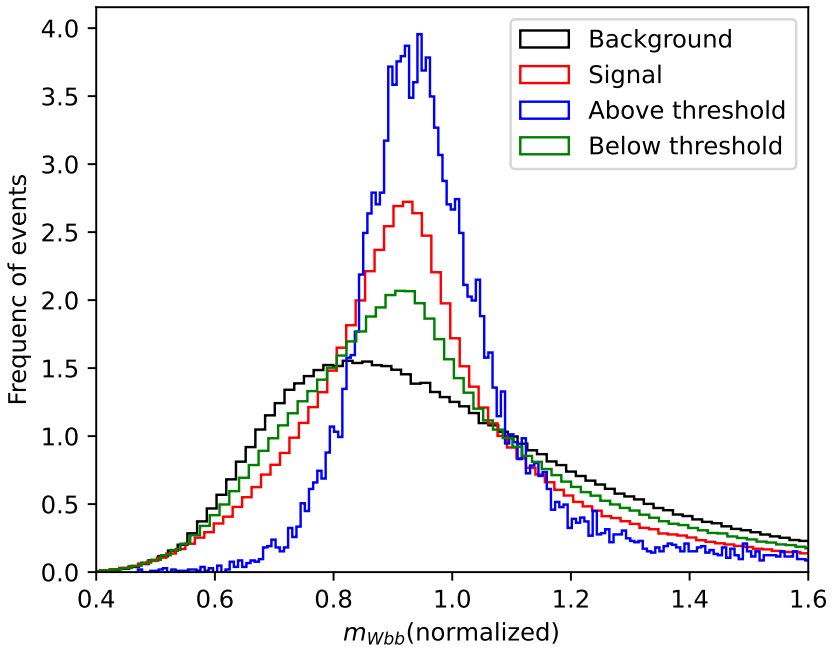


Figure B.3: This figure visualizes the distribution of the 26th feature, the invariant mass m_{Wbb}. The red and black lines are the histograms of the original dataset. We employ MMD-M as a classifier, trained and evaluated using $n_{\text{tr}} = 1.3 \times 10^6$ and $n_{\text{ev}} = n_{\text{opt}} = 2 \times 10^4$ through Algorithm 3. The blue(green) line represents all instances z's whose "witness scores" $f(z; X^{\text{ev}}, Y^{\text{ev}})$'s are larger(smaller) than t_{opt}.

Configuration and Model Architecture

We implement all methods in Python 3.9 and PyTorch 1.13 and run them on an NVIDIA Quadro RTX 8000 GPU.

For all classifier-based methods in this study (SCHE and LBI), we adopt the same architecture as previously proposed in [14]. The classifiers are six-layer neural networks with 300 hidden units in each layer, all employing the tanh activation function. For SCHE, the output layer is a single sigmoid unit and we utilize the binary cross-entropy loss for training. For LBI, the output layer is a linear unit and we utilize the binary cross entropy loss combined with a logit function (which is more numerically stable than simply using a sigmoid layer followed by a cross entropy loss).

For all MMD-based methods (MMD-M, MMD-G, MMD-O, and UME), the networks φ and φ' are both six-layer neural networks with 300 ReLU units in each layer. The feature space, which is the output of the neural network φ, is set to be 100-dimensional. Here UME has the same kernel architecture as MMD-M, and the number of test locations is set to be $J_q = 4096$. For RFM, we adopt the same architecture as in [170], where the kernel is $K_M(x, y) = \exp(-\gamma(x-y)^T M(x-y))$ with a constant γ and a learnable positive semi-definite matrix M. We set $\gamma \equiv 1$.

The neural networks are initialized using the default setting in PyTorch, and the bandwidths σ, σ' are initialized using the *median heuristic* [87]. The parameter τ is initially set to 0.5.

For UME, the witness locations W are initially randomly sampled from the training set. For RFM, the initial M equals the median bandwidth times an identity matrix.

Training

The size of our training set, denoted as n_{tr}, varies from 1.0×10^2 to 1.6×10^6. For a given n_{tr}, we select the first n_{tr} datapoints from each class of the Higgs dataset to form X^{tr} and Y^{tr}, i.e., $|X^{\mathsf{tr}}| = |Y^{\mathsf{tr}}| = n_{\mathsf{tr}}$. Subsequently, we randomly select $n_{\mathsf{validation}} = \min(\sqrt{10n_{\mathsf{tr}}}, 0.1n_{\mathsf{tr}})$ points from each of $X_{\mathsf{tr}}, Y_{\mathsf{tr}}$ to constitute the validation set, while the remainder of $X_{\mathsf{tr}}, Y_{\mathsf{tr}}$ are used for running gradient descent. The optimizer is set to be a minibatch SGD, with a batch size of 1024, a learning rate of 0.001, and a momentum of 0.99. Training is halted once the validation loss stops to decrease for 10 epochs, then we choose the checkpoint (saved for each epoch) with the smallest validation loss thus far as our trained model. Beyond the general setting above, in RFM a batch size of 1024 doesn't work well and instead we use a batch size of $20,000$.

B.8.3 Evaluating the Performance

Evaluating the p-Value with the Methodology of Algorithm 1

We call the "witness score" of an instance $z \in \mathcal{X}$ as

$$f(z; X^{\mathsf{ev}}, Y^{\mathsf{ev}}) = \frac{1}{n_{\mathsf{cal}}} \sum_{i=1}^{n_{\mathsf{cal}}} \left(k(z, Y_i^{\mathsf{ev}}) - k(z, X_i^{\mathsf{ev}})\right). \tag{B.8.1}$$

For a vector of instances $Z = (Z_1, \ldots, Z_m)$, we write

$$f(Z; X^{\mathsf{ev}}, Y^{\mathsf{ev}}) = (f(Z_1; X^{\mathsf{ev}}, Y^{\mathsf{ev}}), \ldots, f(Z_m; X^{\mathsf{ev}}, Y^{\mathsf{ev}})).$$

The testing procedure is summarized in Phases 2, 3 and 4 in Algorithm 2 and Algorithm 3. In the Higgs experiment, we utilize the Gaussian approximation method to determine the p-values when the witness function f is not thresholded, which allows us to reach very small p-values and errors under limited computational resource. In cases where the score function f is thresholded by a value t, using the Binomial distribution as in Algorithm 3 is more precise and also fast enough.

Given a trained kernel K trained on X^{tr} and Y^{tr}, we set $X^{\mathsf{ev}} = X^{\mathsf{tr}}$ and $Y^{\mathsf{ev}} = Y^{\mathsf{tr}}$, and accordingly $n_{\mathsf{ev}} = n_{\mathsf{tr}}$. This results in a more efficient use of data (since we reuse $X^{\mathsf{tr}}, Y^{\mathsf{tr}}$ also as $X^{\mathsf{ev}}, Y^{\mathsf{ev}}$). Then, out of the untouched portion of the data, we randomly choose $n_{\mathsf{cal}} = 20,000$ datapoints from both classes to populate X^{cal} and Y^{cal}, i.e., $|X^{\mathsf{cal}}| = |Y^{\mathsf{cal}}| = n_{\mathsf{cal}} = 20,000$. In addition to the general setting above, for RFM, we need to solve a $2n_{\mathsf{ev}}$-dimensional linear equation during inference, which arises from the inverse matrix in Equation (B.6.3) (solving $K(X^{\mathrm{RFM}}, X^{\mathrm{RFM}})\boldsymbol{u} = (y^{\mathrm{RFM}})^T$ for $\boldsymbol{u} \in \mathbb{R}^{2n_{\mathsf{ev}}}$). So we set $n_{\mathsf{ev}} = \min(n_{\mathsf{tr}}, 10,000)$ that $X_{\mathsf{ev}}, Y_{\mathsf{ev}}$ are randomly sampled from the training set.

In order to compare different benchmarks, we evaluate the expected significance of discovery on a mixture of 1000 backgrounds and 100 signals. For each benchmark and each n_{tr}, we train 10 independent models. Then for each trained model we proceed through the Phases 2, 3 (and

4) in Algorithm 2 and Algorithm 3 by 10 times for 10 different $(X^{\mathsf{ev}}, X^{\mathsf{cal}}, X^{\mathsf{opt}}, Y^{\mathsf{ev}}, Y^{\mathsf{cal}}, Y^{\mathsf{opt}})$. The mean and standard deviation from these 100 runs are reported in Figure B.4.

We also display in Figure B.5 the trade-off b (m, n_{ev}) and (m, n_{tr}) to reach certain levels of significance of discovery in MMD-M. From the bottom left plot, we see that the (averaged) significance is not sensitive to n_{ev} when $\lg n_{\mathsf{ev}}$ is large. So taking $n_{\mathsf{ev}} = 20,000$ is sufficient.

Evaluating the Error of the Test (3.2.4)

We set the parameters to be $\delta = 0.1$ and $\pi = \frac{1}{2}\delta$ in our experiments. As explained Appendix B.7.3, here we no longer take $X^{\mathsf{ev}} = X^{\mathsf{tr}}$. Empirically, taking $X^{\mathsf{ev}} = X^{\mathsf{tr}}$ yields a very bad threshold $\gamma(X^{\mathsf{ev}}, Y^{\mathsf{ev}}, \pi)$.[3] Instead, X^{ev} is sampled from untouched datapoints other than X^{tr}, and the same applies for Y. We still take $n_{\mathsf{ev}} = n_{\mathsf{tr}}$ here, resulting in a total size of $n_{\mathsf{ev}} + n_{\mathsf{tr}} = 2n_{\mathsf{tr}}$. Specifically, when $n_{\mathsf{ev}} \geq 10,000$, computing a $n_{\mathsf{ev}} \times n_{\mathsf{ev}}$ Gram matrix becomes computationally expensive, so we adopt Monte Carlo method to compute $\gamma(X^{\mathsf{ev}}, Y^{\mathsf{ev}}, \pi)$, in which we subsample $10,000$ points from X^{ev} and Y^{ev} to calculate γ and repeat this process 100 times.

Again, we utilize the Gaussian approximation. Recall that the test is to compare $T = \frac{1}{m}\sum_{i=1}^{m} f(Z_i)$ with γ. The type 1 and type 2 error are estimated as

$$\mathrm{CDF}_{\mathcal{N}(0,1)}\left(-\frac{\gamma(X^{\mathsf{ev}}, Y^{\mathsf{ev}}, \pi) - \mathbb{E}[f|H_0]}{\sqrt{\mathrm{var}(f|H_0)/m}}\right)$$

and

$$\mathrm{CDF}_{\mathcal{N}(0,1)}\left(-\frac{\mathbb{E}[f|H_1] - \gamma(X^{\mathsf{ev}}, Y^{\mathsf{ev}}, \pi)}{\sqrt{\mathrm{var}(f|H_1)/m}}\right)$$

for the witness function f, which can be estimated efficiently using the calibration samples $X^{\mathsf{cal}}, Y^{\mathsf{cal}}$.

We consider both the regimes of fixing kernels and varying kernels (training kernel based on n). The results are shown in the top plot in Figure 3.1 and the top plot in Figure B.5. For each point on the plot, we train 30 independent models and test each model 10 times, and report the average of these 300 runs. In both plots, we observe the asymmetric m vs n trade-off.

Algorithm 2 Estimate the significance of discovery of an input Z_{test}, using the original statistic

---

Input: $(X^{\mathsf{tr}}, X^{\mathsf{ev}}, X^{\mathsf{cal}})$, $(Y^{\mathsf{tr}}, Y^{\mathsf{ev}}, Y^{\mathsf{cal}})$; parametrized kernel K_ω; input Z_{test}.

 # Phase 1: Kernel training on X^{tr} and Y^{tr}

$\omega \leftarrow \arg\max_\omega^{\text{optimizer}} \hat{J}(X^{\mathsf{tr}}, Y^{\mathsf{tr}}; K_w)$ *# maximize objective $\widehat{J}(X^{\mathsf{tr}}, Y^{\mathsf{tr}}; K_\omega)$ as in (3.4.1)*

 # Phase 2: Distributional calibration of test statistic

$\text{Scores}^{(0)} \leftarrow f(X^{\mathsf{cal}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}})$ *# $\text{Scores}^{(0)}$ has a length of n_{cal}*

$\text{Scores}^{(1)} \leftarrow f(Y^{\mathsf{cal}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}})$ *# $\text{Scores}^{(1)}$ has a length of n_{cal}*

---

[3] If the kernel $K(\cdot,\cdot) = K_{X^{\mathsf{tr}}, Y^{\mathsf{tr}}}(\cdot,\cdot)$ is independent of $X^{\mathsf{ev}}, Y^{\mathsf{ev}}$, then we have $\gamma(X^{\mathsf{ev}}, Y^{\mathsf{ev}}, \delta/2) \approx \frac{1}{2}\left(\mathbb{E}_{Z\sim P_x}[T(X^{\mathsf{ev}}, Y^{\mathsf{ev}}, Z)] + \mathbb{E}_{Z\sim \delta P_Y + (1-\delta)P_X}[T(X^{\mathsf{ev}}, Y^{\mathsf{ev}}, Z)]\right)$. However this is no longer true if $(X^{\mathsf{tr}}, Y^{\mathsf{tr}})$ and $(X^{\mathsf{ev}}, Y^{\mathsf{ev}})$ intersect.

$\theta_0 \leftarrow \text{mean}(\text{Scores}^{(0)})$ # *estimate* $\mathbb{E}[f(Z)|Z \sim P_X]$
$\theta_1 \leftarrow \text{mean}(\text{Scores}^{(1)})$ # *estimate* $\mathbb{E}[f(Z)|Z \sim P_Y]$
$\sigma_0 \leftarrow \text{std}(\text{Scores}^{(0)})$ # *estimate* $\sqrt{\text{var}[f(Z)|Z \sim P_X]}$
 # Phase 3: Inference with input Z_{test}
$m \leftarrow \text{length}(Z_{\text{test}})$
$T \leftarrow T_f(Z_{\text{test}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}}) = \text{mean}(f(Z_{\text{test}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}}))$ # *compute test statistic*
$Z_{\text{discovery}} \leftarrow \frac{T-\theta_0}{\sigma_0/\sqrt{m}}$
Output: Estimated significance: $Z_{\text{discovery}}$

Algorithm 3 Estimate the significance of discovery of an input Z_{test}, applying the thresholding trick

---

Input: $(X^{\mathsf{tr}}, X^{\mathsf{ev}}, X^{\mathsf{cal}}, X^{\mathsf{opt}})$, $(Y^{\mathsf{tr}}, Y^{\mathsf{ev}}, Y^{\mathsf{cal}}, Y^{\mathsf{opt}})$; parametrized kernel K_ω; input Z_{test}.
 # Phase 1: Kernel training on X^{tr} and Y^{tr}
$\omega \leftarrow \arg\max_\omega^{\text{optimizer}} \hat{J}(X^{\mathsf{tr}}, Y^{\mathsf{tr}}; K_w)$ # *maximize objective $\widehat{J}(X^{\mathsf{tr}}, Y^{\mathsf{tr}}; K_\omega)$ as in (3.4.1)*
 # Phase 2: Find the best threshold
$\text{Scores}^{(0)} \leftarrow f(X^{\mathsf{opt}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}})$
$\text{Scores}^{(1)} \leftarrow f(Y^{\mathsf{opt}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}})$ # *witness function as in (B.8.1)*
for $i = 1, 2, ..., 2n_{\mathsf{opt}}$ **do**
  $t = (\text{Scores}^{(0)} \cup \text{Scores}^{(1)})[i]$
  $\text{TP}, \text{TN} = \text{mean}(\text{Scores}^{(1)} > t), \text{mean}(\text{Scores}^{(0)} < t)$ # *true positive and true negative rate*
  $\text{power}_i = \frac{\text{TP}+\text{TN}-1}{\sqrt{\text{TN}(1-\text{TN})}}$ # *find t to maximize the (estimated) p-value*
end for
$t_{\mathsf{opt}} = (\text{Scores}^{(0)} \cup \text{Scores}^{(1)})[\arg\max_i \text{power}_i]$
 # Phase 3: Distributional calibration of test statistic (under null hypothesis)
$\text{Scores}^{(0)} \leftarrow (f(X^{\mathsf{cal}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}}) > t)$ # *Scores*$^{(0)} \in \{0,1\}^{n_{\mathsf{ev}}}$
$\text{Scores}^{(1)} \leftarrow (f(Y^{\mathsf{cal}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}}) > t)$ # *Scores*$^{(1)} \in \{0,1\}^{n_{\mathsf{ev}}}$
$\theta_0 \leftarrow \text{mean}(\text{Scores}^{(0)})$ # *estimate* $\mathbb{E}[f_t(Z)|Z \sim P_X] \in [0,1]$
$\theta_1 \leftarrow \text{mean}(\text{Scores}^{(1)})$ # *estimate* $\mathbb{E}[f_t(Z)|Z \sim P_Y] \in [0,1]$
 # Phase 4: Inference with input Z_{test}
$m \leftarrow \text{length}(Z_{\text{test}})$
$T \leftarrow T_f(Z_{\text{test}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}}) = \text{mean}(f(Z_{\text{test}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}}) > t)$ # *compute test statistic*
$Z_{\text{discovery}} \leftarrow \text{CDF}^{-1}_{\mathcal{N}(0,1)}(\text{CDF}_{\text{Bin}(m,\theta_0)}(T))$
Output: Estimated significance: $Z_{\text{discovery}}$

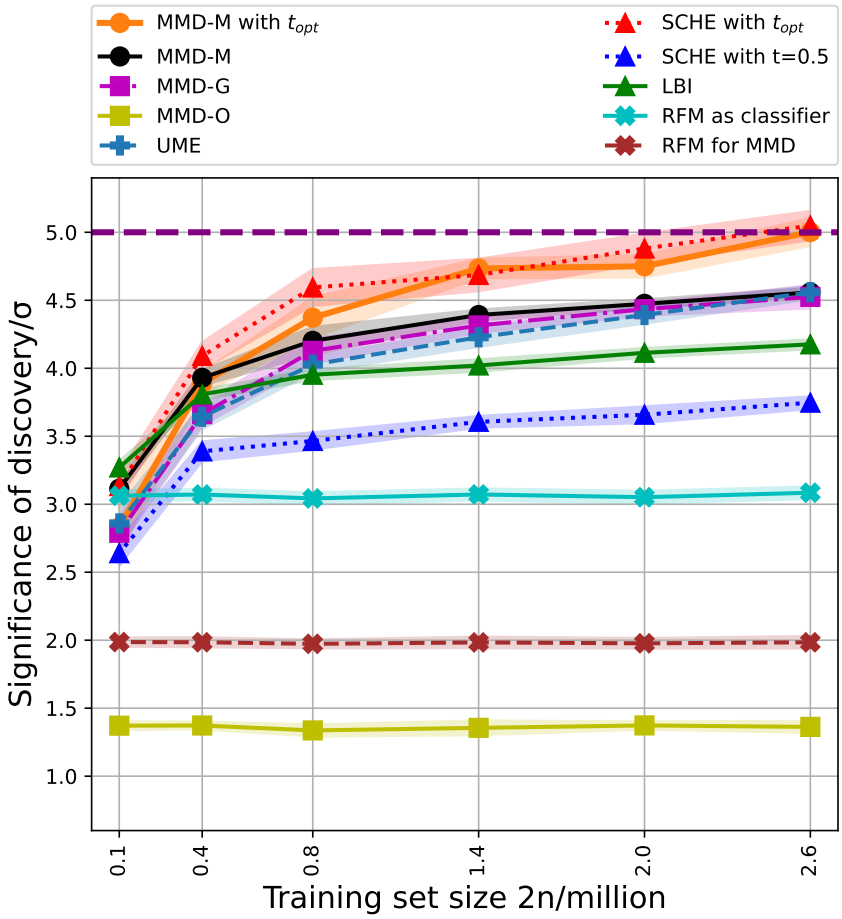


Figure B.4: Complete image of Figure 3.1 in the main text. The mean and standard deviation are calculated based on 100 runs. See Appendix B.8 for details.

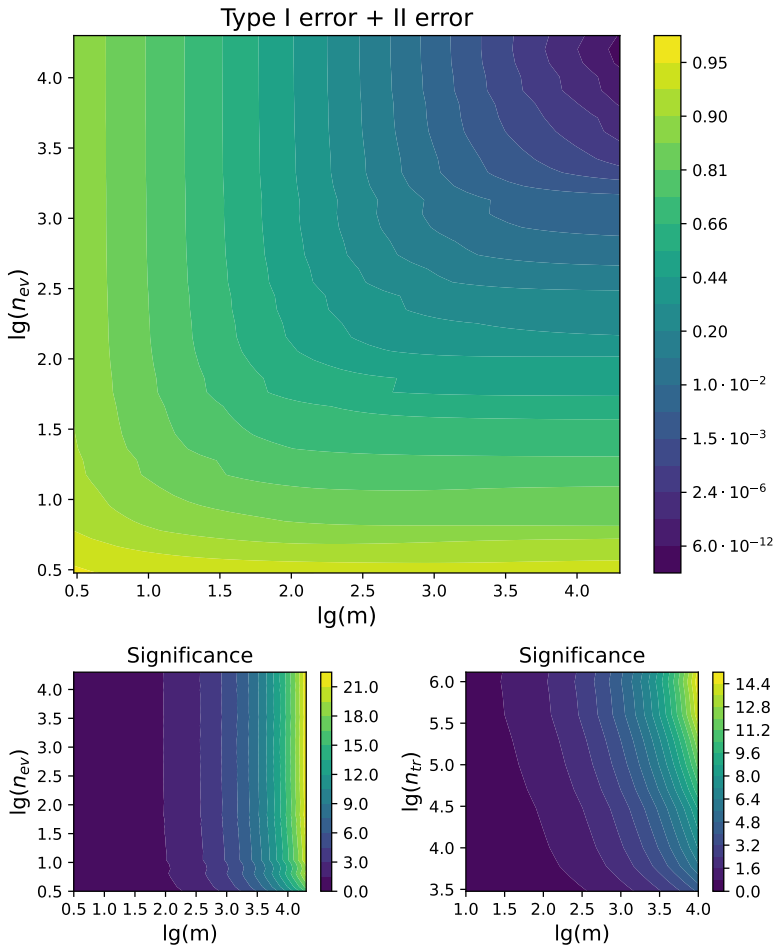


Figure B.5: The top plot displays the (m, n_{ev}) trade-off to reach certain levels of total error using $n_{\mathsf{tr}} = 1.3 \times 10^6$ in MMD-M. The bottom figures show the trade-off of (m, n_{ev}) and (m, n_{tr}) to reach certain level of significance of discovery in MMD-M. In the bottom left figure, we fix $n_{\mathsf{tr}} = 1.3 \times 10^6$. In the bottom right figure, we fix $n_{\mathsf{ev}} = 20,000$. See Appendix B.8 for details.

B.9 Limitations and Future Directions

Finally, we discuss several limitations of our work and raise open questions that we hope will be addressed in future works. From the theoretical side of our arguments, we point out several aspects. First, our upper bound (on the minimax sample complexity) Theorem 3.3.1 has a likely sub-optimal dependence on α, δ. Second, it might be possible to improve our lower bound to a more natural form by replacing $\|\lambda\|_{2,J_\epsilon^\star}$ by $\|\lambda\|_2$ and removing the constraint that the top eigenfunction has to be constant. Third, it remains open to extend our theory to include data-dependent K, as opposed to fixed K.

Empirically, our proposal Algorithm 1 can be inefficient in Phase 2 (prior works such as [143] have used permutation-based arguments for a more efficient estimate), which we adopted due to its simplicity and universality in all benchmarks. Moreover, one might hope

that LFHT/mLFHT can be extended to more complex applications, such as text data or videos. Such questions are important to investigate as a future direction.

Appendix C

Appendix of "Minimax Optimal Testing via Classification"

C.1 Auxiliary Lemmas

We state some auxiliary lemmas which will be used for the proof. We begin with a simple identity for standard normal distributions.

Lemma C.1.1. *Take $a, b \in \mathbb{R}$ and let Z be standard normal. Then*

$$\mathbb{E}\Phi(aZ+b) = \Phi\left(\frac{b}{\sqrt{1+a^2}}\right).$$

Proof. Let Z' be a standard Gaussian independent of Z. Then

$$\mathbb{E}\Phi(aZ+b) = \mathbb{P}(aZ+b \geq Z') = \mathbb{P}\left(\frac{Z'-aZ}{\sqrt{1+a^2}} \leq \frac{b}{\sqrt{1+a^2}}\right) = \Phi\left(\frac{b}{\sqrt{1+a^2}}\right).$$

□

The following lemma is the celebrated result of Gaussian Lipschitz concentration.

Lemma C.1.2 (Lipschitz concentration for Gaussians [201, Theorem 5.2.1])**.** *Let Q be a d-dimensional standard Gaussian and let $f : \mathbb{R}^d \to \mathbb{R}$ be σ-Lipschitz. Then $f(Q)$ is sub-Gaussian with variance proxy σ^2.*

The next lemma states the Chernoff bound for Poisson random variables.

Lemma C.1.3 ([149, Theorem 5.4])**.** *For all $\lambda > 0$ and $x \geq 0$ we have*

$$\mathbb{P}(\mathrm{Poi}(\lambda) - \lambda \geq x) \leq \exp\left(-\frac{x^2}{2(\lambda+x)}\right),$$
$$\mathbb{P}(\mathrm{Poi}(\lambda) - \lambda \leq -x) \leq \exp\left(-\frac{x^2}{2\lambda}\right).$$

The following technical lemma is helpful in establishing the Bernstein concentration in Lemma C.2.1.

Lemma C.1.4. *Let $a \geq 0, p, q \in [0,1]$ and define $\tau = p(1-p) \wedge q(1-q), \nu = p(1-p) \vee q(1-q)$. Then it always holds that*

$$a\sqrt{\frac{\nu}{2}} \leq a\sqrt{\tau} + a^2 + |p-q|.$$

In particular, if $|p-q| \geq a\sqrt{\tau} + a^2$, then

$$4|p-q| \geq a\sqrt{\tau} + a\sqrt{\nu} + a^2.$$

Proof. After rearranging and noting that $1 + 2\sqrt{2} < 4$, it is clear that the first inequality implies the second. Below we prove the first inequality.

Since the claim is invariant under the transformations $(p,q) \mapsto (q,p)$ and $(p,q) \mapsto (1-p, 1-q)$, it suffices to consider the case where $p \leq 1/2$ and $p(1-p) \leq q(1-q)$. It further suffices to consider the case where $p \leq q \leq 1/2$: if not, then $p \leq 1-q \leq 1/2$, and the transformation $(p,q) \mapsto (p, 1-q)$ keeps (τ, ν) invariant while makes $|p-q|$ smaller. The proof is then completed by considering the following two scenarios:

- if $p \geq q/2$, then $\nu = q(1-q) \leq 2p(1-p) = 2\tau$, so $a\sqrt{\nu/2} \leq a\sqrt{\tau}$;
- if $p \leq q/2$, then $2a\sqrt{\nu} \leq a^2 + \nu \leq a^2 + q \leq a^2 + 2(q-p)$.

□

C.2 Omitted Proofs from Section 4.1

C.2.1 Proof of Lemma 4.1.2

Before we prove Lemma 4.1.2, we begin with a technical lemma on the Bernstein concentration of the classifier-accuracy test (4.1.2).

Lemma C.2.1. *Suppose $A_1, \ldots, A_n \overset{iid}{\sim} \mathrm{Ber}(p)$ and $B_1, \ldots, B_m \overset{iid}{\sim} \mathrm{Ber}(q)$. Let $\tau = p(1-p) \wedge q(1-q)$ and define the averages $\bar{A} = \frac{1}{n}\sum_{i=1}^n A_i$ and $\bar{B} = \frac{1}{m}\sum_{j=1}^m B_j$. There exists a universal constant $c > 0$ such that*

$$\mathbb{P}\left(\left|\bar{A} - \bar{B}\right| \leq \frac{1}{2}|p-q| - \frac{1}{2}\sqrt{\frac{c\log(1/\delta)\tau}{n \wedge m}} - \frac{1}{2}\frac{c\log(1/\delta)}{n \wedge m}\right) \leq \delta,$$

$$\mathbb{P}\left(\left|\bar{A} - \bar{B}\right| \geq 2|p-q| + 2\sqrt{\frac{c\log(1/\delta)\tau}{n \wedge m}} + 2\frac{c\log(1/\delta)}{n \wedge m}\right) \leq \delta.$$

Proof. Let $\nu = p(1-p) \vee q(1-q)$. Note that the first inequality is trivially true if

$$|p-q| \leq \sqrt{\frac{c\log(1/\delta)\tau}{n \wedge m}} + \frac{c\log(1/\delta)}{n \wedge m}.$$

Assuming otherwise, by the second statement of Lemma C.1.4, the first probability is upper bounded by

$$\mathbb{P}\left(\left|\bar{A}-\bar{B}\right| \le |p-q| - \frac{5}{8}\sqrt{\frac{c\log(1/\delta)\tau}{n\wedge m}} - \frac{1}{8}\sqrt{\frac{c\log(1/\delta)\nu}{n\wedge m}} - \frac{5}{8}\frac{c\log(1/\delta)}{n\wedge m}\right).$$

By choosing c sufficiently large (independently of p, q, n, m, δ), and applying Bernstein's inequality separately to both $\bar{A}$ and $\bar{B}$, the above probability can be made smaller than δ.

For the second inequality, using the first statement of Lemma C.1.4, it is upper bounded by

$$\mathbb{P}\left(|\bar{A}-\bar{B}| \ge |p-q| + \sqrt{\frac{c\log(1/\delta)\tau}{n\wedge m}} + \frac{1}{\sqrt{2}}\sqrt{\frac{c\log(1/\delta)\nu}{n\wedge m}} + \frac{c\log(1/\delta)}{n\wedge m}\right).$$

Again, taking c sufficiently large (independently of p, q, n, m, δ) and applying Bernstein's inequality separately to both $\bar{A}$ and $\bar{B}$, the above probability can be made smaller than δ. □

Now we proceed to prove Lemma 4.1.2. Using n test samples (X, Y) from both p and q, consider the following classifier-accuracy test: we accept H_0 if

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}(X_i \in S) - \mathbb{1}(Y_i \in S)\right)\right| \le \sqrt{\frac{c\overline{\tau}\log(1/\delta)}{n}} + \frac{c\log(1/\delta)}{n},$$

and reject H_0 otherwise. Here $c > 0$ is a large absolute constant, and we note that the threshold only relies on the knowledge of $\overline{\tau}$ in addition to (n, δ).

To analyze the type-I and type-II errors, first assume that H_0 holds. Since $\mathrm{sep}(S) = 0$ under H_0, the second statement of Lemma C.2.1 implies that we accept H_0 with probability at least $1 - \delta/2$ if $c > 0$ is large enough. If H_1 holds, with probability at least $1 - \delta/2$, by the first statement of Lemma C.2.1 we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}(X_i \in S) - \mathbb{1}(Y_i \in S)\right)\right| \ge |\underline{\mathrm{sep}}| - \left(\sqrt{\frac{c\overline{\tau}\log(1/\delta)}{n}} + \frac{c\log(1/\delta)}{n}\right).$$

By the lower bound of n assumed in Lemma 4.1.2, in this case we will reject H_0, as desired.

C.2.2 Proof of Proposition 4.1.3

Lemma C.2.2. *Let μ be a non-negative measure on some space $\mathcal{X}$ and let $a, b : \mathcal{X} \to \mathbb{R}_+$ such that $\int a(x)\mathrm{d}\mu(x) > 0$ and $b(x) = 0$ only if $a(x) = 0$. Then*

$$\inf_{x\in\mathrm{spt}(\mu)}\left(\frac{a(x)}{b(x)}\right) \le \frac{\int a(x)\mathrm{d}\mu(x)}{\int b(x)\mathrm{d}\mu(x)} \le \sup_{x\in\mathrm{spt}(\mu)}\left(\frac{a(x)}{b(x)}\right).$$

Proof. Defining $0/0 = 1$, we have

$$\begin{aligned}\int a(x)\mathrm{d}\mu(x) &= \int \frac{a(x)}{b(x)}b(x)\mathrm{d}\mu(x)\\ &\le \sup_{x\in\mathrm{spt}(\mu)}\left(\frac{a(x)}{b(x)}\right)\int b(x)\mathrm{d}\mu(x).\end{aligned}$$

The other direction follows analogously. □

Proof of Proposition 4.1.3. Let p, q be the densities of $\mathbb{P}, \mathbb{Q}$ with respect to a common dominating measure, and let $E =: \{x : p(x) > q(x)\}$ so that $\mathsf{TV}(\mathbb{P}, \mathbb{Q}) = \mathbb{P}(E) - \mathbb{Q}(E) > 0$. Assume without loss of generality that $\mathbb{P}(E) + \mathbb{Q}(E) \geq 1$. Given $t \in [0, 1]$ define $E_t =: \{x : \frac{p(x)-q(x)}{p(x)+q(x)} \geq t\}$, so that the map $t \mapsto \mathbb{P}(E_t) + \mathbb{Q}(E_t)$ is non-increasing and left-continuous. Note that $E_0 = E$ while $E_1 = \varnothing$, so that $t^\star = \max\{t \in [0, 1] : \mathbb{P}(E_t) + \mathbb{Q}(E_t) \geq 1\}$ exists. Now choose the randomized classifier $\mathcal{C}$ as follows:

$$\mathcal{C}(x) = \begin{cases} 0 & \text{if } x \in E_{(t^\star)^+}, \\ 1 & \text{if } x \notin E_{t^\star}, \\ \mathrm{Ber}(r) & \text{if } x \in E_{t^\star} - E_{(t^\star)^+}, \end{cases}$$

where $E_{(t^\star)^+} = \cap_{t > t^\star} E_t \subseteq E_{t^\star}$, and

$$r := \frac{1 - \mathbb{P}(E_{(t^\star)^+}) - \mathbb{Q}(E_{(t^\star)^+})}{\mathbb{P}(E_{t^\star}) + \mathbb{Q}(E_{t^\star}) - \mathbb{P}(E_{(t^\star)^+}) - \mathbb{Q}(E_{(t^\star)^+})} \in [0, 1].$$

This classifier is balanced, as

$$\begin{aligned} &\mathbb{P}(\mathcal{C}(X) = 0) + \mathbb{Q}(\mathcal{C}(X) = 0) \\ &= \mathbb{P}(E_{(t^\star)^+}) + \mathbb{Q}(E_{(t^\star)^+}) + r(\mathbb{P}(E_{t^\star}) + \mathbb{Q}(E_{t^\star}) - \mathbb{P}(E_{(t^\star)^+}) - \mathbb{Q}(E_{(t^\star)^+})) \\ &= 1. \end{aligned}$$

For $t \in [0, 1]$ define

$$f(t) =: \begin{cases} (\mathbb{P}(E_t) - \mathbb{Q}(E_t))/(\mathbb{P}(E_t) + \mathbb{Q}(E_t)) & \text{if } \mathbb{P}(E_t) + \mathbb{Q}(E_t) > 0, \\ 1 & \text{otherwise.} \end{cases}$$

Let $0 \leq t \leq s \leq 1$, we show that $f(t) \leq f(s)$. Without loss of generality assume that $f(s) < 1$ and that $\mathbb{P}(E_s \backslash E_t) + \mathbb{Q}(E_s \backslash E_t) > 0$. Notice that $f(t) \leq f(s)$ if and only if

$$\frac{\int_{E_t \backslash E_s} (p(x) - q(x)) \mathrm{d}x}{\int_{E_t \backslash E_s} (p(x) + q(x)) \mathrm{d}x} \overset{!}{\leq} \frac{\int_{E_s} (p(x) - q(x)) \mathrm{d}x}{\int_{E_s} (p(x) + q(x)) \mathrm{d}x}.$$

However, the above inequality follows from Lemma C.2.2. Thus, it holds that

$$\frac{\mathbb{P}(\mathcal{C}(X) = 0) - \mathbb{Q}(\mathcal{C}(X) = 0)}{\mathbb{P}(\mathcal{C}(X) = 0) + \mathbb{Q}(\mathcal{C}(X) = 0)} \geq f(t^\star) \geq f(0) = \frac{\mathbb{P}(E) - \mathbb{Q}(E)}{\mathbb{P}(E) + \mathbb{Q}(E)}.$$

Plugging in $\mathbb{P}(\mathcal{C}(X) = 0) + \mathbb{Q}(\mathcal{C}(X) = 0) = 1$ and $\mathbb{P}(E) + \mathbb{Q}(E) \leq 2$ yields the result.

To show tightness, one can consider $p(x) = \mathbb{1}_{[0,1]}$, $q(x) = (1 + \epsilon)\mathbb{1}_{[0,1/(1+\epsilon)]}$, $\mathcal{C}(x) = \mathbb{1}_{x \in (1/(2+\epsilon), 1]}$, and let $\epsilon \to 0^+$. □

C.3 Omitted Proofs from Section 4.3

C.3.1 Useful Lemmas

Before we present the formal proofs, this section summarizes some useful lemmas on the expected value and sub-Gaussian concentration of the separation.

Lemma C.3.1. *Let $\mu \geq \lambda \geq 0$ and $X \sim \mathrm{Poi}(\mu), Y \sim \mathrm{Poi}(\lambda)$. Then*

$$\mathbb{P}(X > Y) + \frac{1}{2}\mathbb{P}(X = Y) - \frac{1}{2} \geq c\left(\frac{\mu - \lambda}{\sqrt{\lambda + 1}} \wedge 1\right)$$

holds, where $c > 0$ is a universal constant.

Proof. For $t \in [\lambda, \mu]$ define the function

$$f(t) = \mathbb{P}(\mathrm{Poi}(t) > Y) + \frac{1}{2}\mathbb{P}(\mathrm{Poi}(t) = Y).$$

Clearly $f(\lambda) = \frac{1}{2}$. We have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}(\mathrm{Poi}(t) > Y) = -\mathbb{P}(\mathrm{Poi}(t) > Y) + \mathbb{P}(\mathrm{Poi}(t) > Y - 1) = \mathbb{P}(\mathrm{Poi}(t) = Y).$$

Similarly we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}(\mathrm{Poi}(t) = Y) = -\mathbb{P}(\mathrm{Poi}(t) = Y) + \mathbb{P}(\mathrm{Poi}(t) = Y - 1).$$

Thus, we obtain

$$f'(t) = \frac{1}{2}\mathbb{E}\left[\mathbb{P}(\mathrm{Poi}(t) \in \{Y - 1, Y\})\right].$$

Next we prove the following inequality: if y is a non-negative integer with $|y - t| \leq 8\sqrt{t}$, then

$$\mathbb{P}(\mathrm{Poi}(t) = y) = \Omega\left(\frac{1}{\sqrt{t + 1}}\right). \tag{C.3.1}$$

To prove (C.3.1), we distinguish three scenarios:

1. If $t < 1/100$, then the only non-negative integer y with $|y - t| \leq 8\sqrt{t}$ is $y = 0$. Therefore $\mathbb{P}(\mathrm{Poi}(t) = y) = e^{-t} = \Omega(1)$.

2. If $1/100 \leq t \leq 100$, then $0 \leq y \leq 180$. In this case,

$$\mathbb{P}(\mathrm{Poi}(t) = y) \geq \min_{1/100 \leq t \leq 100} \min_{0 \leq y \leq 180} \mathbb{P}(\mathrm{Poi}(t) = y) = \Omega(1).$$

3. If $t > 100$, then for $t - 8\sqrt{t} \leq y_1 \leq y_2 \leq t + 8\sqrt{t}$, we have

$$\frac{\mathbb{P}(\mathrm{Poi}(t) = y_1)}{\mathbb{P}(\mathrm{Poi}(t) = y_2)} = t^{y_2 - y_1}\frac{y_2!}{y_1!} = \prod_{y = y_1 + 1}^{y_2} \frac{t}{y} = \left(1 \pm \mathcal{O}(t^{-1/2})\right)^{\mathcal{O}(16\sqrt{t})} = \Theta(1).$$

   In the above we have used that $|t/y - 1| = \mathcal{O}(t^{-1/2})$ for all $y \in [y_1, y_2]$, and $y_2 - y_1 \leq 16\sqrt{t}$. Consequently,

$$\mathbb{P}(\mathrm{Poi}(t) = y) = \Omega\left(\frac{\mathbb{P}(|\mathrm{Poi}(t) - t| \leq 8\sqrt{t})}{16\sqrt{t}}\right) = \Omega\left(\frac{1}{\sqrt{t}}\right),$$

   where the last step is due to Chebyshev's inequality.

Now we apply (C.3.1) to prove Lemma C.3.1. We first show that for non-negative integer y,

$$\{|y-\lambda| \le 2\sqrt{\lambda}\} \wedge \{\sqrt{\lambda} \le \sqrt{t} \le \sqrt{\lambda}+1\} \Longrightarrow \{|y-t| \le 8\sqrt{t}\}. \tag{C.3.2}$$

In fact, if $\sqrt{\lambda} < \sqrt{2}-1$, then the LHS of (C.3.2) implies that $y=0$ and $t<2$, thus (C.3.2) holds. If $\sqrt{\lambda} \ge \sqrt{2}-1$, then the LHS of (C.3.2) implies that

$$|y-t| \le |y-\lambda| + (t-\lambda) \le 2\sqrt{\lambda} + (2\sqrt{\lambda}+1) < 8\sqrt{\lambda} \le 8\sqrt{t},$$

and (C.3.2) holds as well. Next, by (C.3.1) and (C.3.2), as well as Chebyshev's inequality $\mathbb{P}(|Y-\lambda| \le 2\sqrt{\lambda}) \ge \frac{3}{4}$, we have

$$\begin{aligned}
f'(t) &\ge \frac{3}{8} \min_{y \ge 0 : |y-\lambda| \le 2\sqrt{\lambda}} \mathbb{P}(\mathrm{Poi}(t)=y) \\
&\ge \frac{3}{8} \mathbb{1}\{\sqrt{\lambda} \le \sqrt{t} \le \sqrt{\lambda}+1\} \cdot \min_{|y-t| \le 8\sqrt{t}} \mathbb{P}(\mathrm{Poi}(t)=y) \\
&= \Omega\left(\frac{\mathbb{1}\{\sqrt{\lambda} \le \sqrt{t} \le \sqrt{\lambda}+1\}}{\sqrt{t+1}}\right) = \Omega\left(\frac{\mathbb{1}\{\sqrt{\lambda} \le \sqrt{t} \le \sqrt{\lambda}+1\}}{\sqrt{\lambda+1}}\right).
\end{aligned}$$

Finally, for some absolute constant $c>0$ it holds that

$$f(\mu) - f(\lambda) = \int_\lambda^\mu f'(t)\mathrm{d}t \ge c \int_\lambda^\mu \frac{\mathbb{1}\{\sqrt{\lambda} \le \sqrt{t} \le \sqrt{\lambda}+1\}}{\sqrt{\lambda+1}}\mathrm{d}t \ge c\left(\frac{\mu-\lambda}{\sqrt{\lambda+1}} \wedge 1\right),$$

which is the statement of the lemma. □

Lemma C.3.2. *For any $D \subseteq [k]$, each of $\mathrm{sep}(\hat{S}_s(D)), s \in \{>, <, 1/2\}$ is sub-Gaussian with variance proxy σ^2 which can be bounded as*

$$\sigma^2 \lesssim \sum_{i \in D} (p_i - q_i)^2 \wedge \frac{p_i+q_i}{n} = \mathcal{O}\left(\frac{1}{n}\right),$$

with universal hidden constants.

Proof. Using standard tail bounds of the Poisson distribution (Lemma C.1.3) we have for any $i \in D$ with $p_i > q_i$,

$$\begin{aligned}
&\mathbb{P}(i \in \hat{S}_<(D)) \le \mathbb{P}(i \notin \hat{S}_{1/2}(D)) \le \mathbb{P}(i \notin \hat{S}_>(D)) \\
&= \mathbb{P}(\mathrm{Poi}(np_i) \le \mathrm{Poi}(nq_i)) \\
&\le \mathbb{P}\left(\mathrm{Poi}(np_i) - np_i \le -\frac{1}{2}n(p_i-q_i)\right) + \mathbb{P}\left(\mathrm{Poi}(nq_i) - nq_i > \frac{1}{2}n(p_i-q_i)\right) \\
&\le 2\exp\left(-c\frac{n(p_i-q_i)^2}{p_i+q_i}\right)
\end{aligned}$$

for some universal $c > 0$. Similarly, if $i \in D$ with $p_i \le q_i$ we get

$$\begin{aligned}\mathbb{P}(i \in \hat{S}_{>}(D)) \le \mathbb{P}(i \in \hat{S}_{1/2}(D)) &\le \mathbb{P}(i \notin \hat{S}_{<}(D)) \\ &= \mathbb{P}(\mathrm{Poi}(np_i) \ge \mathrm{Poi}(nq_i)) \le 2\exp\left(-c\frac{n(p_i-q_i)^2}{p_i+q_i}\right).\end{aligned}$$

Using these estimates we turn to bounding the moment generating function of $\mathrm{sep}(\hat{S}_s)$ for $s \in \{>, <, 1/2\}$. Before doing so, recall [34, Theorem 2.1] that the best-possible sub-Gaussian variance proxy $\sigma^2_{\mathsf{opt}}(\mu)$ of the $\mathrm{Ber}(\mu)$ distribution satisfies

$$\sigma^2_{\mathsf{opt}}(\mu) = \frac{\frac{1}{2}-\mu}{\log\left(\frac{1}{\mu}-1\right)},$$

where the values for $\mu \in \{0, \frac{1}{2}, 1\}$ should be understood as the limit of the above expression (resulting in $\sigma^2_{\mathsf{opt}} = 0, \frac{1}{4}, 0$ respectively). Notice also that $\mu \mapsto \sigma^2_{\mathsf{opt}}(\mu)$ is increasing on $[0, \frac{1}{2}]$ and decreasing on $[\frac{1}{2}, 1]$, and

$$\sigma^2_{\mathsf{opt}}(\mu) \le \begin{cases} \frac{2}{\log(2/\mu)} & \text{if } 0 < \mu < 1/4, \\ 1/4 & \text{if } 1/4 \le \mu \le 3/4, \\ \frac{2}{\log(2/(1-\mu))} & \text{if } 3/4 < \mu < 1. \end{cases}$$

Let $T \subseteq D$ denote the subset of indices given by

$$T = \left\{i \in D : 2\exp\left(-c\frac{n(p_i-q_i)^2}{p_i+q_i}\right) \ge \frac{1}{4}\right\} = \left\{i \in D : (p_i-q_i)^2 \le \frac{p_i+q_i}{n}\frac{\log(8)}{c}\right\}.$$

Now, for any $s \in \{>, <, 1/2\}$, the sub-Gaussian variance proxy σ_s^2 of $\mathrm{sep}(\hat{S}_s) - \mathbb{E}\,\mathrm{sep}(\hat{S}_s) = \sum_{i\in D}(p_i - q_i)(\mathbb{1}\{i \in \hat{S}_s\} - \mathbb{P}(i \in \hat{S}_s)))$ is at most

$$\sigma_s^2 \le \sum_{i\in T}\frac{(p_i-q_i)^2}{4} + \sum_{i\in D\setminus T}(p_i-q_i)^2 \cdot \frac{2(p_i+q_i)}{cn(p_i-q_i)^2} \lesssim \sum_{i\in D}(p_i-q_i)^2 \wedge \frac{p_i+q_i}{n},$$

where the second step used the definition of T. In particular, since $\sum_{i\in D}(p_i+q_i)/n \le 2/n$, the above expression is always upper bounded by $\mathcal{O}(1/n)$. □

C.3.2 Proof of Proposition 4.3.1

By Lemma C.3.1, we have

$$\begin{aligned}\mathbb{E}\,\mathrm{sep}(\hat{S}_{1/2}) &= \sum_{i\in[k]}\mathbb{P}(i \in \hat{S}_{1/2})(p_i - q_i) \\ &= \sum_{i\in[k]}(\mathbb{P}(i \in \hat{S}_{1/2}) - \frac{1}{2})(p_i - q_i) \\ &\gtrsim \sum_{i\in[k]}\left(\frac{n|p_i-q_i|}{\sqrt{n(p_i\wedge q_i)+1}} \wedge 1\right)|p_i - q_i| \\ &\ge \min_{G\subseteq[k]}\left\{\sum_{i\in G}\frac{n(p_i-q_i)^2}{\sqrt{n(q_i\wedge p_i)+1}} + \sum_{i\notin G}|p_i-q_i|\right\}.\end{aligned}$$

Applying the Cauchy-Schwarz inequality twice, we can bound the first term above by

$$\sum_{i\in G}\frac{n(p_i-q_i)^2}{\sqrt{n(q_i+p_i)+1}}\geq\frac{n\left(\sum_{i\in G}|p_i-q_i|\right)^2}{\sum_{i\in G}\sqrt{n(q_i+p_i)+1}}\geq\frac{n\left(\sum_{i\in G}|p_i-q_i|\right)^2}{\sqrt{2nk+k^2}}.$$

Therefore, we get the lower bound

$$\mathbb{E}\,\mathrm{sep}(\hat{S}_{1/2})\gtrsim\min_{0\leq\epsilon_1\leq\epsilon}\left\{\frac{n\epsilon_1^2}{\sqrt{k(n+k)}}+\epsilon-\epsilon_1\right\}=\begin{cases}\frac{\epsilon^2}{\lambda} & \text{if } \epsilon<\frac{\lambda}{2}\\ \epsilon-\frac{\lambda}{4}\geq\frac{\epsilon}{2} & \text{if } \epsilon\geq\frac{\lambda}{2}\end{cases}\gtrsim\epsilon^2\left(\frac{1}{\epsilon}\wedge\sqrt{\frac{n}{k}}\wedge\frac{n}{k}\right)$$

where $\lambda=\frac{\sqrt{k(n+k)}}{n}\asymp\sqrt{\frac{k}{n}}\vee\frac{k}{n}$.

By Lemma C.3.2 we know that $\mathrm{sep}(\hat{S}_{1/2})$ is sub-Gaussian with variance proxy $\mathcal{O}(1/n)$, which implies that $|\mathrm{sep}(\hat{S}_{1/2})|\gtrsim\epsilon^2(\frac{1}{\epsilon}\wedge\sqrt{\frac{n}{k}}\wedge\frac{n}{k})$ with probability at least $1-\delta$, provided that

$$\epsilon^2\left(\frac{1}{\epsilon}\wedge\sqrt{\frac{n}{k}}\wedge\frac{n}{k}\right)\gtrsim\sqrt{\frac{\log(1/\delta)}{n}}.$$

The above rearranges to $n\gtrsim n_{\mathsf{TS}}(\epsilon,\delta,\mathcal{P}_{\mathsf{D}})$.

C.3.3 Proof of Proposition 4.3.2

A direct computation gives

$$\begin{aligned}2\mathbb{E}\,\mathrm{sep}(\hat{S}_{>})&=2\sum_{i=1}^{3k}(p_i-q_i)\mathbb{P}(i\in\hat{S}_{>})\\&=-\mathbb{P}\left(\mathrm{Poi}\left(\frac{n}{2k}\right)>\mathrm{Poi}\left(\frac{n}{k}\right)\right)+1-e^{-n/(4k)}\\&\leq-(1-e^{-n/(2k)})e^{-n/k}+1-e^{-n/(4k)}\\&=-e^{-n/k}+e^{-3n/(2k)}+1-e^{-n/(4k)}\leq 0,\end{aligned}$$

for $\exp(-n/(4k))\gtrapprox 0.86$. Rearranging, this gives the sufficient condition $n/k\leq 0.6$.

C.3.4 Proof of Proposition 4.3.3

Similar to the proof of Proposition 4.3.1, we have by Lemma C.3.1 that

$$\begin{aligned}\mathbb{E}\,\mathrm{sep}(\hat{S}_{1/2}(D))&=\sum_{i\in D}(p_i-q_i)\mathbb{P}(i\in\hat{S}_{1/2}(D))\geq c\mathcal{E}(D)+\frac{1}{2}\{p(D)-q(D)\}\\-\mathbb{E}\,\mathrm{sep}(D\setminus\hat{S}_{1/2}(D))&=\sum_{i\in D}(q_i-p_i)\mathbb{P}(i\notin\hat{S}_{1/2}(D))\geq c\mathcal{E}(D)+\frac{1}{2}\{q(D)-p(D)\}\end{aligned}$$

where $c>0$ is universal and $\mathcal{E}(D)=\sum_{i\in D}\frac{n|p_i-q_i|^2}{\sqrt{n(p_i\wedge q_i)+1}}\wedge|p_i-q_i|$. Therefore,

$$\begin{aligned}&\mathbb{E}\left[\mathrm{sep}(\hat{S}_{>}(D))-\mathrm{sep}(\hat{S}_{<}(D))\right]\\&\quad=\mathbb{E}\left[\mathrm{sep}(\hat{S}_{1/2}(D))-\mathrm{sep}(D\setminus\hat{S}_{1/2}(D))\right]\geq 2c\mathcal{E}(D).\end{aligned}\tag{C.3.3}$$

The bound on the sub-Gaussian variance proxy follows directly from Lemma C.3.2.

C.3.5 Proof of Corollary 4.3.4

By a two-fold sample splitting, suppose that we have independent held out samples $(\tilde{X}, \tilde{Y})$ identical in distribution to (X, Y). In the sequel we will use samples (X, Y) to construct two separating sets, and use samples $(\tilde{X}, \tilde{Y})$ to make a choice between them.

Let the sets $\hat{S}_> =: \hat{S}_>([k]), \hat{S}_< =: \hat{S}_<([k])$ be constructed using X, Y. By Proposition 4.3.1 and 4.3.3, we have

$$|\mathbb{E}\,\mathrm{sep}(\hat{S}_>)| \vee |\mathbb{E}\,\mathrm{sep}(\hat{S}_<)| \gtrsim \epsilon^2 \left(\frac{1}{\epsilon} \wedge \sqrt{\frac{n}{k}} \wedge \frac{n}{k}\right),$$
$$\sigma^2(\hat{S}_>) + \sigma^2(\hat{S}_<) \lesssim \sum_{i\in[k]} \frac{p_i + q_i}{n} \lesssim \frac{1}{k \vee n},$$

where the last step have used that $p_i + q_i \lesssim 1/k$ in $\mathcal{P}_{\mathsf{Db}}$. Going forward, we assume that

$$\epsilon^2 \left(\frac{1}{\epsilon} \wedge \sqrt{\frac{n}{k}} \wedge \frac{n}{k}\right) \gtrsim \sqrt{\frac{\log(1/\delta)}{k \vee n}},$$

which rearranges to $n \gtrsim n_{\mathsf{GoF}}(\epsilon, \delta, \mathcal{P}_{\mathsf{D}})$. Consequently, this ensures that $|\,\mathrm{sep}(\hat{S}_>)| \vee |\,\mathrm{sep}(\hat{S}_<)| \gtrsim \epsilon^2 \left(\frac{1}{\epsilon} \wedge \sqrt{\frac{n}{k}} \wedge \frac{n}{k}\right)$ with probability $1 - \mathcal{O}(\delta)$. Moreover, as $n \gtrsim \log(1/\delta)$, with probability at least $1-\delta$ we have $\mathrm{Poi}(n) \le 2n$ (cf. Lemma C.1.3). Under this event, one has $|\hat{S}_>| \vee |\hat{S}_<| \le 2n$, and

$$\tau(\hat{S}_>) \vee \tau(\hat{S}_<) \lesssim \frac{|\hat{S}_>| \vee |\hat{S}_<|}{k} \wedge 1 \le \frac{2n}{k} \wedge 1.$$

Next we make a choice between $\hat{S}_>$ and $\hat{S}_<$ based on held out samples $(\tilde{X}, \tilde{Y})$. Let $\hat{p}, \hat{q}$ denote the empirical pmfs constructed using $\tilde{X}, \tilde{Y}$ respectively. For any set $A \subseteq [k]$ write $\widehat{\mathrm{sep}}(A) = \hat{p}(A) - \hat{q}(A)$. We define our final estimator to be

$$\hat{S} = \begin{cases} \hat{S}_> & \text{if } |\widehat{\mathrm{sep}}(\hat{S}_>)| \ge |\widehat{\mathrm{sep}}(\hat{S}_<)|, \\ \hat{S}_< & \text{otherwise.} \end{cases}$$

Clearly $\tau(\hat{S}) \le \tau(\hat{S}_>) \vee \tau(\hat{S}_<) \lesssim 1 \wedge (n/k)$. To show the high-probability separation of $\hat{S}$, note that by Lemma C.2.1, it holds with probability at least $1 - \mathcal{O}(\delta)$ that

$$\begin{aligned}
|\,\mathrm{sep}(\hat{S})| &\ge \frac{1}{2}|\widehat{\mathrm{sep}}(\hat{S})| - \mathcal{O}\left(\sqrt{\frac{\tau(\hat{S})\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n}\right) \\
&= \frac{1}{2}|\widehat{\mathrm{sep}}(\hat{S}_>)| \vee |\widehat{\mathrm{sep}}(\hat{S}_<)| - \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n \vee k}} + \frac{\log(1/\delta)}{n}\right) \\
&\ge \frac{1}{4}|\,\mathrm{sep}(\hat{S}_>)| \vee |\,\mathrm{sep}(\hat{S}_<)| - \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n \vee k}} + \frac{\log(1/\delta)}{n}\right) \\
&= \Omega\left(\epsilon^2 \left(\frac{1}{\epsilon} \wedge \sqrt{\frac{n}{k}} \wedge \frac{n}{k}\right)\right) - \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n \vee k}} + \frac{\log(1/\delta)}{n}\right).
\end{aligned}$$

Here the first term always dominates the second as long as $n \gtrsim n_{\mathsf{GoF}}(\epsilon, \delta, \mathcal{P}_{\mathsf{D}})$.

C.3.6 Proof of Proposition 4.3.6

Similar to the proof of Corollary 4.3.4, we apply a two-fold sample splitting to obtain n independent held out samples $(\tilde{X}, \tilde{Y})$. In the sequel we construct $2(\ell+2)$ candidate separating sets from (X, Y), and make a choice among them using held out samples $(\tilde{X}, \tilde{Y})$.

The construction of the $2(\ell+2)$ separating sets is simple: for each $j \in \{0, 1, \cdots, \ell+1\}$, we construct two sets $\hat{S}_>(D_j)$ and $\hat{S}_<(D_j)$. The following lemma summarizes some properties of these separating sets. Recall that we assume that $t = k \wedge (c_0 m/\log(1/\delta)) > n$ so that $\ell = \lceil \log_2(t/n) \rceil \geq 1$.

Lemma C.3.3. *Fix any $j \in \{0, 1, \cdots, \ell+1\}$, and let $\epsilon_j = \sum_{i \in D_j} |p_i - q_i|$. With probability at least $1 - \delta$, the following statements hold:*

1. *if $j = 0$, then*

$$\left|\mathrm{sep}(\hat{S}_>(D_0))\right| \vee \left|\mathrm{sep}(\hat{S}_<(D_0))\right| \gtrsim E_0 - \mathcal{O}\left(\sqrt{\frac{E_0 \log(1/\delta)}{n}}\right),$$

   where

$$E_0 = \sum_{i \in D_0} n|p_i - q_i|^2 \wedge |p_i - q_i| \gtrsim \frac{n\epsilon_0^2}{k} =: \tilde{E}_0(\epsilon_0).$$

2. *if $j \in [\ell]$, then*

$$\left|\mathrm{sep}(\hat{S}_>(D_j))\right| \vee \left|\mathrm{sep}(\hat{S}_<(D_j))\right| \gtrsim E_j - \mathcal{O}\left(\sqrt{\frac{E_j \log(1/\delta)}{n}}\right),$$

   where

$$E_j = \sum_{i \in D_j} n|p_i - q_i|^2 \wedge |p_i - q_i| \gtrsim \frac{n\epsilon_j^2}{\sqrt{kt/2^j}} =: \tilde{E}_j(\epsilon_j).$$

3. *if $j = \ell + 1$, then*

$$\left|\mathrm{sep}(\hat{S}_>(D_{\ell+1}))\right| \vee \left|\mathrm{sep}(\hat{S}_<(D_{\ell+1}))\right| \gtrsim E_{\ell+1} - \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

   where

$$E_{\ell+1} = \sum_{i \in D_{\ell+1}} \frac{n|p_i - q_i|^2}{\sqrt{nq_i}} \wedge |p_i - q_i| \gtrsim \sqrt{\frac{n}{k}}\epsilon_{\ell+1}^2 =: \tilde{E}_{\ell+1}(\epsilon_{\ell+1}).$$

Proof. We prove the above statements separately.

1. Case I: $j = 0$. By Proposition 4.3.3, it holds that

$$\mathbb{E}[\mathrm{sep}(\hat{S}_>(D_0)) - \mathrm{sep}(\hat{S}_<(D_0))] \gtrsim \sum_{i \in D_0} n|p_i - q_i|^2 \wedge |p_i - q_i| = E_0,$$

where we have used Lemma 4.3.5 that $q_i \le 2/t \le 2/n$ for all $i \in D_0$. Moreover,

$$\begin{aligned}&\sigma^2(\mathrm{sep}(\hat{S}_>(D_0))) \vee \sigma^2(\mathrm{sep}(\hat{S}_<(D_0)))\\ &\lesssim \sum_{i \in D_0} |p_i - q_i|^2 \wedge \frac{p_i + q_i}{n} \lesssim \sum_{i \in D_0} \frac{1}{n}\left(n|p_i - q_i|^2 \wedge |p_i - q_i|\right) = \frac{E_0}{n},\end{aligned}$$

where the last inequality is due to the following deterministic inequality: if $q \le 2/n$, then

$$|p - q|^2 \wedge \frac{p + q}{n} \lesssim \frac{1}{n}\left(n|p - q|^2 \wedge |p - q|\right).$$

The proof of the above deterministic inequality is based on two cases:

- if $p \le 3/n$, then $|p - q|^2 \lesssim |p - q|^2 \wedge (|p - q|/n)$;
- if $p > 3/n$, then $p + q \lesssim n|p - q|^2 \wedge |p - q|$.

Consequently, we have the first statement. For the second statement, similar to the proof of Proposition 4.3.1 we have

$$E_0 \ge \min_{\epsilon_0' \in [0, \epsilon_0]} \left(\frac{n(\epsilon_0')^2}{k} + \epsilon_0 - \epsilon_0'\right) \gtrsim \epsilon_0^2 \left(\frac{1}{\epsilon_0} \wedge \frac{n}{k}\right) \asymp \frac{n\epsilon_0^2}{k}.$$

2. Case II: $j \in [\ell]$. By Proposition 4.3.3 and Lemma 4.3.5 we have

$$\mathbb{E}[\mathrm{sep}(\hat{S}_>(D_j)) - \mathrm{sep}(\hat{S}_<(D_j))] \gtrsim \sum_{i \in D_j} n(p_i - q_i)^2 \wedge |p_i - q_i| = E_j.$$

Similar to Case I, we have

$$\sigma^2(\mathrm{sep}(\hat{S}_>(D_j))) \vee \sigma^2(\mathrm{sep}(\hat{S}_<(D_j))) \lesssim \sum_{i \in D_j} |p_i - q_i|^2 \wedge \frac{p_i + q_i}{n} \lesssim \frac{E_j}{n},$$

and the first statement follows.

For the second statement, note that $|D_j| \le t/2^{j-1} = \mathcal{O}(\sqrt{kt/2^j})$ by Lemma 4.3.5. Therefore,

$$E_j \ge \min_{\epsilon_j' \in [0, \epsilon_j]} \left(\frac{n(\epsilon_j')^2}{|D_j|} + \epsilon_j - \epsilon_j'\right) \gtrsim \epsilon_j^2 \left(\frac{1}{\epsilon_j} \wedge \frac{n}{\sqrt{kt/2^j}}\right) \asymp \frac{n\epsilon_j^2}{\sqrt{kt/2^j}}.$$

3. Case III: $j = \ell + 1$. By Proposition 4.3.3 and Lemma 4.3.5, we have

$$\mathbb{E}[\mathrm{sep}(\hat{S}_>(D_{\ell+1})) - \mathrm{sep}(\hat{S}_<(D_{\ell+1}))] \gtrsim \sum_{i \in D_{\ell+1}} \frac{n(p_i - q_i)^2}{\sqrt{nq_i}} \wedge |p_i - q_i| = E_{\ell+1}.$$

The first statement then follows from Lemma C.3.2. The second statement then follows from

$$E_{\ell+1} \geq \min_{\epsilon'_{\ell+1} \in [0, \epsilon_{\ell+1}]} \left(\frac{n(\epsilon'_{\ell+1})^2}{\sqrt{nk}} + \epsilon_{\ell+1} - \epsilon'_{\ell+1} \right) \gtrsim \epsilon_{\ell+1}^2 \left(\frac{1}{\epsilon_{\ell+1}} \wedge \sqrt{\frac{n}{k}} \right) \asymp \sqrt{\frac{n}{k}} \epsilon_{\ell+1}^2.$$

The proof is complete. □

Based on Lemma C.3.3, we are about to describe how we choose from the sets $\{\hat{S}_>(D_j), \hat{S}_<(D_j)\}_{j=0}^{\ell+1}$. Similar to the proof of Corollary 4.3.4, using the held out samples $(\tilde{X}, \tilde{Y})$, we can obtain the empirical estimates $\widehat{\mathrm{sep}}(\hat{S}_s(D_j))$ for all $s \in \{>, <\}$ and $j \in \{0, 1, \cdots, \ell + 1\}$. With a small absolute constant $c_1 > 0$ and $\tilde{E}_j$ as defined in Lemma C.3.3, the selection rule is as follows: if there is some $s \in \{>, <\}$ and $j \in \{0, 1, \cdots, \ell + 1\}$ such that

$$|\widehat{\mathrm{sep}}(\hat{S}_s(D_j))| \geq c_1 \tilde{E}_j(\epsilon/(\ell + 2)),$$

then choose $\hat{S} = \hat{S}_s(D_j)$; if there is no such pair (s, j), choose an arbitrary $\hat{S}$.

We first show that with probability at least $1 - \mathcal{O}(k\delta)$, such a pair (s, j) exists. Since $\|p - q\|_1 \geq \epsilon$, there must exist some $j \in \{0, 1, \cdots, \ell + 1\}$ such that $\epsilon_j \geq \epsilon/(\ell + 2)$. As long as

$$n \geq c_2 n_{\mathsf{GoF}}(\epsilon/\ell, \delta, \mathcal{P}_{\mathsf{D}})$$

for a large constant $c_2 > 0$, one can check via Lemma C.3.3 that $|\mathrm{sep}(\hat{S}_>(D_j))| \vee |\mathrm{sep}(\hat{S}_<(D_j))| \geq 4c_1 \tilde{E}_j(\epsilon/(\ell + 2))$ for a small enough universal constant $c_1 > 0$. Assuming that $n \gtrsim \log(1/\delta)$, we have $\tau(\hat{S}_>(D_j)) \vee \tau(\hat{S}_<(D_j)) = \mathcal{O}(n2^j/t)$ with probability $1 - \mathcal{O}(\delta)$ due to Poisson concentration (Lemma C.1.3). On this event, it holds with probability at least $1 - \delta$ that (cf. Lemma C.2.1)

$$|\widehat{\mathrm{sep}}(\hat{S}_>(D_j))| \vee |\widehat{\mathrm{sep}}(\hat{S}_<(D_j))| \geq 2c_1 \tilde{E}_j(\epsilon/(\ell + 2)) - \mathcal{O}\left(\sqrt{\frac{2^j \log(1/\delta)}{t}} + \frac{\log(1/\delta)}{n} \right),$$

which is at least $c_1 \tilde{E}_j(\epsilon/(\ell + 2))$ as long as

$$n\sqrt{\frac{t}{k}} \asymp n\sqrt{1 \wedge \frac{m}{\log(1/\delta)k}} \geq c_3 n_{\mathsf{GoF}}(\epsilon/\ell, \delta, \mathcal{P}_{\mathsf{D}}) \tag{C.3.4}$$

for some large $c_3 > 0$. Therefore, provided (C.3.4) holds, the desired pair (j, s) exists with probability $1 - \mathcal{O}(k\delta)$ due to a union bound.

Conversely, if $|\widehat{\mathrm{sep}}(\hat{S}_s(D_j))| \geq c_1 \tilde{E}_j(\epsilon/(\ell + 2))$ holds for some (s, j), the true separation $|\mathrm{sep}(\hat{S}_s(D_j))|$ is at least of the same order as well. Indeed, Lemma C.2.1 shows that

$$|\mathrm{sep}(\hat{S}_s(D_j))| \geq \frac{1}{2} |\widehat{\mathrm{sep}}(\hat{S}_s(D_j))| - \mathcal{O}\left(\sqrt{\frac{2^j \log(1/\delta)}{t}} + \frac{\log(1/\delta)}{n} \right),$$

which is at least $c_1 E_j(\epsilon/(\ell + 2))/4$ as long as (C.3.4) holds. This completes the proof.

C.3.7 Proof of Proposition 4.3.8

The statement of Proposition 4.3.8 follows immediately from the following lemma.

Lemma C.3.4. *Let* $\mathrm{sep}(\hat{S}) =: \mu_{\theta^X}(\hat{S}) - \mu_{\theta^Y}(\hat{S})$. *There exist universal constants* $c_i > 0, i \in [5]$ *such that for* $J = \lfloor c_1 \epsilon^{-1/s} \rfloor$ *we have*

$$\mathbb{E}[\mathrm{sep}(\hat{S})] + \frac{c_2}{\sqrt{n}} \geq \frac{c_3 \epsilon^2}{\epsilon + \sqrt{J/n}}$$
$$\mathbb{P}\left(\left| \mathrm{sep}(\hat{S}) - \mathbb{E}\,\mathrm{sep}(\hat{S}) \right| \geq t + \frac{c_4}{\sqrt{n}} \right) \leq 2\exp(-c_5 n t^2)$$

for all $t \geq 0$.

Proof. Write $\|\cdot\|, \langle \cdot, \cdot \rangle$ for the ℓ^2 norm/inner product restricted to the first J coordinates. Notice that given $\hat{\theta}^X$ and $\hat{\theta}^Y$, $T(\theta)$ is simply a Gaussian random variable with $\mathbb{E}T(\theta) = \|\hat{\theta}^Y - \theta\|^2 - \|\hat{\theta}^X - \theta\|^2$ and $\mathrm{var}(T) = 4\|\hat{\theta}^X - \hat{\theta}^Y\|^2$. Define the vectors

$$U = \{\hat{\theta}^X_j - \hat{\theta}^Y_j\}_{j=1}^J$$
$$V = \{\hat{\theta}^X_j + \hat{\theta}^Y_j\}_{j=1}^J.$$

Note that they are independent, jointly Gaussain with variance $2I_J/n$ and means equal to the first J coordinates of $\theta^X \mp \theta^Y$ respectively. Let Φ be the cdf of the standard Gaussian and $\phi = \Phi'$ be its density. The separation can be written as

$$\mathrm{sep}(\hat{S}) = f(\theta^X) - f(\theta^Y),$$

where

$$f(\theta) = \Phi\left(\frac{\|\hat{\theta}^Y - \theta\|^2 - \|\hat{\theta}^X - \theta\|^2}{2\|\hat{\theta}^X - \hat{\theta}^Y\|} \right) = \Phi\left(-\frac{1}{2}\left\langle V, \frac{U}{\|U\|} \right\rangle + \left\langle \theta, \frac{U}{\|U\|} \right\rangle \right). \tag{C.3.5}$$

We focus on proving the desired tail bound first. To make the dependence on the variables explicit, write $g(U, V) = f(\theta^X) - f(\theta^Y)$ for the separation. Given U, V is a $\mathcal{N}(\theta^X + \theta^Y, 2I_j/n)$ random variable. Differentiating g and using that ϕ is $1/\sqrt{2\pi e}$-Lipschitz we have

$$\begin{aligned}
\|\nabla_V g(U,V)\| &= \Big\| -\frac{1}{2}\frac{U}{\|U\|}\Bigg(\phi\left(-\frac{1}{2}\left\langle V, \frac{U}{\|U\|} \right\rangle + \left\langle \theta^X, \frac{U}{\|U\|} \right\rangle \right) \\
&\qquad - \phi\left(-\frac{1}{2}\left\langle V, \frac{U}{\|U\|} \right\rangle + \left\langle \theta^Y, \frac{U}{\|U\|} \right\rangle \right) \Bigg) \Big\| \\
&\leq \frac{1}{\sqrt{8\pi e}} \left| \left\langle \theta^X - \theta^Y, \frac{U}{\|U\|} \right\rangle \right| \\
&\leq \frac{C_{\mathsf{G}}}{\sqrt{8\pi e}}.
\end{aligned}$$

By Lipschitz concentration of the Gaussian distribution (Lemma C.1.2) we conclude that $g - \mathbb{E}[g|U]$ is sub-Gaussian with variance proxy $C_{\mathsf{G}}^2/(4\pi e n)$. Next we study the concentration of $\mathbb{E}[g|U]$. To this end, note that

$$-\frac{1}{2}\left\langle V, \frac{U}{\|U\|}\right\rangle + \left\langle \theta, \frac{U}{\|U\|}\right\rangle \Bigg| U \sim \mathcal{N}\left(\left\langle \theta - \frac{1}{2}(\theta^X + \theta^Y), \frac{U}{\|U\|}\right\rangle, \frac{1}{2n}\right).$$

Thus, using the independence of U and V and Lemma C.1.1 we obtain

$$\begin{aligned}\mathbb{E}\left[g(U,V)|U\right] &= \mathbb{E}\left[f(\theta^X) - f(\theta^Y)|U\right] \\ &= \Phi\left(\frac{W}{\sqrt{4+2/n}}\right) - \Phi\left(-\frac{W}{\sqrt{4+2/n}}\right),\end{aligned}$$

where we write $W =: \left\langle \theta^X - \theta^Y, \frac{U}{\|U\|}\right\rangle$. Let $\tilde{\Phi} = \Phi(\cdot/\sqrt{4+2/n})$ to ease notation. Once again by Lipschitzness of Φ, we obtain for every $t \geq 0$ that

$$\begin{aligned}\mathbb{P}\left(\left|\tilde{\Phi}(W) - \mathbb{E}\tilde{\Phi}(W)\right| \geq t\right) &\leq \mathbb{P}\left(\left|\tilde{\Phi}(W) - \tilde{\Phi}(\mathbb{E}W)\right| \geq t - \|\tilde{\Phi}\|_{\mathsf{Lip}}\sqrt{\mathrm{var}(W)}\right) \\ &\leq \mathbb{P}\left(|W - \mathbb{E}W| \geq \frac{t}{\|\tilde{\Phi}\|_{\mathsf{Lip}}} - \sqrt{\mathrm{var}(W)}\right),\end{aligned}$$

and an analogous inequality can be obtained for $-W$. The last ingredient is showing that W concentrates well.

Lemma C.3.5. *W is sub-Gaussian with variance proxy $1/(2n)$.*

Proof of Lemma C.3.5. To simplify notation, let $\tau = \theta^X - \theta^Y$, $\sigma^2 = 1/(2n)$ and let Q be a zero-mean identity-covariance Gaussian random vector so that

$$W \overset{d}{=} \left\langle \tau, \frac{\tau + \sigma Q}{\|\tau + \sigma Q\|}\right\rangle.$$

We have

$$\begin{aligned}\left\langle \tau, \frac{\tau+\sigma Q}{\|\tau+\sigma Q\|}\right\rangle = &\underbrace{\left\langle \frac{\tau}{\mathbb{E}\|\tau+\sigma Q\|}, \frac{\tau+\sigma Q}{\|\tau+\sigma Q\|}\right\rangle}_{|\cdot|\leq 1 \text{ almost surely}} \underbrace{(\|\tau+\sigma Q\| - \mathbb{E}\|\tau+\sigma Q\|)}_{\sigma^2 \text{ sub-Gaussian}} \\ &+ \underbrace{\sigma\left\langle \frac{\tau}{\mathbb{E}\|\tau+\sigma Q\|}, Q\right\rangle}_{\sigma^2 \text{ sub-Gaussian}},\end{aligned}$$

where we use that $\mathbb{E}\|\tau + \sigma Q\| \geq \|\tau\|$ by Jensen's inequality, and apply Lemma C.1.2 twice. Overall, this implies that W is sub-Gaussian with variance proxy $\sigma^2 = 1/(2n)$ as required. □

Recall that we have decomposed the separation as follows:

$$\mathrm{sep}(\hat{S}) - \mathbb{E}\,\mathrm{sep}(\hat{S}) = \underbrace{g - \mathbb{E}[g|U]}_{\mathcal{O}(1/n) \text{ sub-Gaussian}} + \underbrace{\tilde{\Phi}(W) - \tilde{\Phi}(-W) - \mathbb{E}[\tilde{\Phi}(W) - \tilde{\Phi}(-W)]}_{\mathcal{O}(1/n) \text{ sub-Gaussian tails beyond } \mathcal{O}(1/\sqrt{n})},$$

which completes the proof.

Let us turn to calculating the expected separation. We have already seen that

$$\mathbb{E}\,\mathrm{sep}(\hat{S}) = \mathbb{E}\left[\tilde{\Phi}(W) - \tilde{\Phi}(-W)\right].$$

Again by Lipschitzness we have $|\mathbb{E}\tilde{\Phi}(W) - \tilde{\Phi}(\mathbb{E}W)| \le \|\tilde{\Phi}\|_{\mathsf{Lip}}\mathbb{E}|W - \mathbb{E}W| \lesssim 1/\sqrt{n}$ by Lemma C.3.5. Thus, we see that

$$\mathbb{E}\,\mathrm{sep}(\hat{S}) + \Omega\left(\frac{1}{\sqrt{n}}\right) \ge \tilde{\Phi}(\mathbb{E}W) - \tilde{\Phi}(-\mathbb{E}W),$$

where the implied constant is universal. To simplify notation, let $\tau = \theta^X - \theta^Y$, $\sigma^2 = 1/(2n)$ and let Q be a standard normal random variable. Looking at $\mathbb{E}W$ we have

$$\mathbb{E}W = \mathbb{E}\left\langle \tau, \frac{\tau + \sigma Q}{\|\tau + \sigma Q\|}\right\rangle = \frac{1}{\sigma}\mathbb{E}\left\langle \tau, \nabla_Q \|\tau + \sigma Q\|\right\rangle = \frac{1}{\sigma}\mathbb{E}\left[\langle \tau, Q\rangle \|\tau + \sigma Q\|\right]$$

by Stein's identity. By the rotational invariance of the Gaussian distribution, the above is equal to

$$\begin{aligned}
\mathbb{E}W &= \frac{\|\tau\|}{\sigma}\mathbb{E}\left[Q_1\sqrt{(\|\tau\| + \sigma Q_1)^2 + \cdots + \sigma^2 Q_J^2}\right] \\
&= \frac{\|\tau\|}{\sigma}\mathbb{E}\left[Q_1\sqrt{(\|\tau\| + \sigma Q_1)^2 + \cdots + \sigma^2 Q_J^2} - Q_1\sqrt{\|\tau\|^2 + \sigma^2 Q_1^2 + \cdots + \sigma^2 Q_J^2}\right] \\
&= 2\|\tau\|^2\mathbb{E}\left[\frac{Q_1^2}{\sqrt{(\|\tau\| + \sigma Q_1)^2 + \cdots + \sigma^2 Q_J^2} + \sqrt{\|\tau\|^2 + \sigma^2 Q_1^2 + \cdots + \sigma^2 Q_J^2}}\right].
\end{aligned}$$

By the Cauchy-Schwarz inequality we have

$$(\mathbb{E}|Q_1|)^2 \lesssim \mathbb{E}\left[\frac{Q_1^2}{\sqrt{(\|\tau\| + \sigma Q_1)^2 + \cdots + \sigma^2 Q_J^2} + \sqrt{\|\tau\|^2 + \sigma^2 Q_1^2 + \cdots + \sigma^2 Q_J^2}}\right]\left(\|\tau\| + \sigma\sqrt{J}\right).$$

Plugging into our expression for $\mathbb{E}W$ this yields

$$\mathbb{E}W \gtrsim \frac{\|\tau\|^2}{\|\tau\| + \sigma\sqrt{J}}.$$

To clarify notation, let us now write $\|\cdot\|_J$ for the ℓ^2-norm restricted to the first J coordinates. Taking $J = c\epsilon^{-1/s}$ it holds that

$$\|\tau\|_J^2 = \|\tau\|^2 - \sum_{j>J}\tau_j^2 \ge \|\tau\|^2 - J^{-2s}\sum_{j>J}\tau_j^2 j^{2s} = \|\tau\|^2 - c^{-2s}\epsilon^2\|\tau\|_s^2.$$

Since $\|\tau\|_s \lesssim 1$ and $\|\tau\| \ge \epsilon$ by assumption, we see that for large enough universal constant c we have $\|\tau\|_J \ge \epsilon/2$. Since the map $x \mapsto x^2/(x + c)$ is increasing for $x, c > 0$ it follows that

$$\mathbb{E}W \gtrsim \frac{\epsilon^2}{\epsilon + \sqrt{J/n}}$$

for a universal implied constant. By the inequality $\Phi(x) - \Phi(-x) \ge x/2$ for $x \in [0, 1]$ we obtain

$$\tilde{\Phi}(\mathbb{E}W) - \tilde{\Phi}(-\mathbb{E}W) \ge 1 \wedge \mathbb{E}W/2,$$

which completes the proof. □

C.4 Lower bounds

Recall the notation of Section 4.2.1. Given two hypotheses H_0, H_1, our aim is to lower bound the minimum achievable worst-case error. To this end, we use the following standard fact:

$$\min_{\psi} \max_{i=0,1} \sup_{P \in H_i} \mathbb{P}_{S\sim P}(\psi(S) \neq i) \geq \frac{1}{2}\big(1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1)\big), \tag{C.4.1}$$

where P_0, P_1 are any random probability distributions with $\mathbb{P}(P_i \in H_i) = 1$ and $\mathbb{E}P_i$ denote the corresponding mixtures and TV denotes the total variation distance. Hence, deriving a lower bound of order δ on the minimax error reduces to the problem of finding mixtures $\mathbb{E}P_i$ such that $1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) = \Omega(\delta)$. To this end we utilize standard inequalities between divergences.

Lemma C.4.1 ([168])**.** *For any probability measures $\mathbb{P}, \mathbb{Q}$ the inequalities*

$$1 - \mathsf{TV}(\mathbb{P}, \mathbb{Q}) \geq \frac{1}{2} e^{-\mathsf{KL}(\mathbb{P}\|\mathbb{Q})} \geq \frac{1}{2(1 + \chi^2(\mathbb{P}\|\mathbb{Q}))}$$

hold, where KL and χ^2 denote the Kullback-Leibler and χ^2 divergence respectively.

Many of our lower bounds will follow from reduction to prior work.

C.4.1 Lower bounds for $\mathcal{P}_{\mathsf{Db}}$

In [77] the authors gave the construction of distributions $p_{\eta,\epsilon}, p_0 \in \mathcal{P}_{\mathsf{Db}}(k, 2)$ (originally due to Paninski) for a mixing parameter η such that $\mathsf{TV}(p_{\eta,\epsilon}, p_0) = \epsilon \asymp \sqrt{\mathsf{KL}(p_{\eta,\epsilon}, p_0)}$ for all η, where the implied constant is universal. They further showed that

$$\chi^2(\mathbb{E}_\eta p_{\eta,\epsilon}^{\otimes n}, p_0^{\otimes n}) \leq \exp\left(c\frac{n^2\epsilon^4}{k}\right) - 1 \tag{C.4.2}$$

and

$$\chi^2\Big(\mathbb{E}_\eta\big[p_0^{\otimes n} \otimes p_{\epsilon,\eta}^{\otimes(n+m)}\big] \Big\| \mathbb{E}_\eta\big[p_0^{\otimes n} \otimes p_{\epsilon,\eta}^{\otimes n} \otimes p_0^{\otimes m}\big]\Big) \leq \exp\left(c\frac{m(n+m)\epsilon^4}{k}\right) - 1 \tag{C.4.3}$$

for a universal $c > 0$.

Remark 22. *More precisely, (C.4.3) can be extracted from [77] using the chain rule for χ^2 (as opposed to KL).*

Lower bound for TS and GoF

Take $P_0 = p_0^{\otimes 2n}$ and $P_1 = p_{\epsilon,\eta_0}^{\otimes n} \otimes p_0^{\otimes n}$ in (C.4.1) for a fixed η_0. Then, by Lemma C.4.1 and the data-processing inequality we have

$$1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) \geq \frac{1}{2}\exp(-n\mathsf{KL}(p_{\epsilon,\eta}\|p_0)) \geq \frac{1}{2}\exp(-cn\epsilon^2) \overset{!}{=} \Omega(\delta)$$

for a universal $c > 0$. This shows that $\mathsf{GoF}, \mathsf{TS}$ are impossible at total error δ unless $n \gtrsim \log(1/\delta)/\epsilon^2$, which gives the first term of our lower bound.

For the second term, consider the random measures $P_0 = p_0^{\otimes 2n}$ and $P_1 = p_0^{\otimes n} \otimes p_{\epsilon,\eta}^{\otimes n}$ in (C.4.1). Then using (C.4.2) and Lemma C.4.1 we have

$$\begin{aligned} 1 - \mathsf{TV}(\mathbb{E}P_1, \mathbb{E}P_0) &\geq \frac{1}{2} \frac{1}{1 + \chi^2(\mathbb{E}P_1 \| \mathbb{E}P_0)} \\ &\geq \frac{1}{2} \exp\left(-c \frac{n^2\epsilon^4}{k}\right) \stackrel{!}{=} \Omega(\delta). \end{aligned}$$

Therefore, TS is impossible unless $n \gtrsim \sqrt{k \log(1/\delta)}/\epsilon^2$, which yields the second term of our lower bound.

Lower bound for LFHT

The necessity of $m \gtrsim \log(1/\delta)/\epsilon^2$ and $n \gtrsim \sqrt{k \log(1/\delta)}/\epsilon^2$ follows as for TS above. Taking $P_0 = p_0^{\otimes n} \otimes p_{\epsilon,\eta}^{\otimes n} \otimes p_0^{\otimes m}$ and $P_1 = p_0^{\otimes n} \otimes p_{\epsilon,\eta}^{\otimes (n+m)}$ in (C.4.1), using (C.4.3) and Lemma C.4.1 we obtain the inequality

$$\begin{aligned} 1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) &\geq \frac{1}{2} \frac{1}{1 + \chi^2(\mathbb{E}P_1 \| \mathbb{E}P_0)} \\ &\geq \frac{1}{2} \exp\left(-c \frac{m(m+n)\epsilon^4}{k}\right) \stackrel{!}{=} \Omega(\delta). \end{aligned}$$

Therefore, LFHT is impossible with error $\mathcal{O}(\delta)$ unless $mn \gtrsim k \log(1/\delta)/\epsilon^4$ (note that the m^2-term is never active), which completes the lower bound proof.

C.4.2 Lower bounds for $\mathcal{P}_{\mathsf{H}}$

We don't provide the details because they are entirely analogous to Section C.4.1 and rely on classical constructions that can be found in [77].

C.4.3 Lower bounds for $\mathcal{P}_{\mathsf{G}}$

Given a vector $\eta \in \{\pm 1\}^{\mathbb{N}}$ define the measure

$$\mathbb{P}_\eta = \bigotimes_{j=1}^{\infty} \left\{ \begin{array}{ll} \mathcal{N}(\eta_j c_1 \epsilon^{\frac{2s+1}{2s}}, 1) & \text{if } 1 \leq j \leq c_2 \epsilon^{-1/s}, \\ \mathcal{N}(0,1) & \text{otherwise.} \end{array} \right\}$$

Let $\eta_1, \eta_2, \ldots$ be iid uniform signs in $\{\pm 1\}$, and γ_η be the mean vector of $\mathbb{P}_\eta$. Writing $\|\cdot\|_s$ for the Sobolev-norm of smoothness s and $\|\cdot\|$ for the Euclidean norm, we see that for any η

$$\begin{aligned} \|\gamma_\eta\|_s^2 &= \sum_{j=1}^{\infty} j^{2s} \gamma_{\eta j}^2 = \sum_{j=1}^{c_2 \epsilon^{-1/s}} j^{2s} c_1^2 \epsilon^{\frac{2s+1}{s}} \leq c_1^2 \epsilon^{\frac{2s+1}{s}} \left(2 c_2 \epsilon^{-1/s}\right)^{2s+1} \asymp c_1^2 c_2^{2s+1}, \\ \|\gamma_\eta\|^2 &= \sum_{j=1}^{\infty} \gamma_{\eta j}^2 = c_1^2 \epsilon^{\frac{2s+1}{s}} c_2 \epsilon^{-1/s} \asymp c_1^2 c_2 \epsilon^2. \end{aligned}$$

Then for any $C_{\mathsf{G}} > 0$ we can choose c_1, c_2 independently of ϵ such that $\mathbb{P}_0, \mathbb{P}_\eta \in \mathcal{P}_{\mathsf{G}}(s, C_{\mathsf{G}})$ almost surely and $\|\gamma_\eta\| = 10\epsilon$. Then for $\epsilon \le 1/10$ we know that

$$\mathsf{TV}(\mathbb{P}_0, \mathbb{P}_\eta) = 2\Phi\left(\frac{\|\gamma_\eta\|}{2}\right) - 1 \ge \epsilon.$$

Lower bounds for GoF and TS

Take $P_0 = \mathbb{P}_0^{\otimes 2n}$ and $P_1 = \mathbb{P}_{\mathbb{1}}^{\otimes n} \otimes \mathbb{P}_0^{\otimes n}$. Then

$$\mathsf{KL}(P_0 \| P_1) = n\mathsf{KL}(\mathbb{P}_0 \| \mathbb{P}_{\mathbb{1}}) = nc_2\epsilon^{-1/s}\frac{(c_1\epsilon^{\frac{2s+1}{2s}} - 0)^2}{2} \asymp n\epsilon^2.$$

Using Lemma C.4.1 this gives us

$$1 - \mathsf{TV}(P_0, P_1) \gtrsim \exp(-\mathsf{KL}(P_0 \| P_1)) = \exp(-\Theta(n\epsilon^2)) \overset{!}{=} \Omega(\delta).$$

By (C.4.1) we know then that $n \gtrsim \log(1/\delta)/\epsilon^2$ is necessary for both GoF and TS over $\mathcal{P}_{\mathsf{G}}$.

To get the second term in the minimax sample complexity consider the construction $P_0 = \mathbb{P}_0^{\otimes 2n}$ and $P_1 = \mathbb{P}_\eta^{\otimes n} \otimes \mathbb{P}_0^{\otimes n}$ where η is a uniformly random vector of signs. Writing $\omega = c_1\epsilon^{\frac{2s+1}{2s}}$ note that

$$\mathbb{E}\mathbb{P}_\eta^{\otimes n} = \bigotimes_{j=1}^{c_2\epsilon^{-1/s}} \left(\frac{1}{2}\mathcal{N}(\omega, 1)^{\otimes n} + \frac{1}{2}\mathcal{N}(-\omega, 1)^{\otimes n}\right).$$

From here we can compute

$$\begin{aligned}
\mathsf{KL}(P_0 \| \mathbb{E}P_1) &\asymp \epsilon^{-1/s}\mathsf{KL}\left(\mathcal{N}(0,1)^{\otimes n} \middle\| \frac{1}{2}\mathcal{N}(\omega, 1)^{\otimes n} + \frac{1}{2}\mathcal{N}(-\omega, 1)^{\otimes n}\right) \\
&\asymp \epsilon^{-1/s}\left(\frac{n}{2}\omega^2 - \mathbb{E}_{X\sim\mathcal{N}(0,I_n)} \log\cosh\left(\omega\sum_{i=1}^n X_i\right)\right) \\
&\le \frac{\epsilon^{-1/s}}{4}n^2\omega^4 \asymp n^2\epsilon^{\frac{4s+1}{s}},
\end{aligned}$$

where we used the inequality $\log\cosh(x) \ge \frac{x^2}{2} - \frac{x^4}{12}$ for all $x \in \mathbb{R}$. Thus, using Lemma C.4.1,

$$1 - \mathsf{TV}(P_0 \| \mathbb{E}P_1) \gtrsim \exp(-\mathsf{KL}(P_0 \| \mathbb{E}\mathbb{P}_1)) \ge \exp(-\Theta(n^2\epsilon^{\frac{4s+1}{s}})) \overset{!}{=} \Omega(\delta).$$

By (C.4.1) we know then that $n \gtrsim \sqrt{\log(1/\delta)}/\epsilon^{\frac{2s+1/2}{s}}$ is necessary for both GoF and TS over $\mathcal{P}_{\mathsf{G}}$.

Lower bounds for LFHT

If $m \ge n$, from the GoF lower bound $n \gtrsim n_{\mathsf{GoF}}$ we conclude that $mn \gtrsim n_{\mathsf{GoF}}^2$, as desired. Therefore, throughout this section we assume that $m < n$.

Let $P_0 = \mathbb{P}_\eta^{\otimes n} \otimes \mathbb{P}_0^{\otimes n} \otimes \mathbb{P}_\eta^m$ and $P_1 = \mathbb{P}_\eta^{\otimes n} \otimes \mathbb{P}_0^{\otimes n} \otimes \mathbb{P}_0^{\otimes m}$, where η is a uniformly random vector of signs. Once again, we define $\omega = c_1 \epsilon^{\frac{2s+1}{2s}}$. We follow a proof similar to the cases $\mathcal{P}_{\mathsf{Db}}, \mathcal{P}_{\mathsf{H}}$ in [77]. We use the data processing inequality, the chain rule and tensorization of χ^2:

$$\begin{aligned}\chi^2(\mathbb{E}P_0 \| \mathbb{E}P_1) &= \chi^2(\mathbb{E}\mathbb{P}_\eta^{\otimes(n+m)} \| \mathbb{E}\mathbb{P}_\eta^{\otimes n} \otimes \mathbb{P}_0^{\otimes m}) \\ &= \left(\mathbb{E}_{X_1} \mathbb{E}_{\eta_1|X} \mathbb{E}_{\eta_1'|X_1} \int_{\mathbb{R}^m} \frac{\exp\left(-\frac{1}{2}\sum_{j=1}^m \{(z_j - \eta_1\omega)^2 + (z_j - \eta_1'\omega)^2\}\right)}{(2\pi)^{m/2}\exp(-\frac{1}{2}\sum_{j=1}^m z_j^2)} \mathrm{d}z \right)^{c_2\epsilon^{-1/s}} - 1,\end{aligned}$$

where $X_1 \sim (\frac{1}{2}\mathcal{N}(\omega, 1/n) + \frac{1}{2}\mathcal{N}(-\omega, 1/n))$ and $\eta_1, \eta_1'|X_1$ are iid scalar signs from the posterior $p(\cdot|X_1)$, with joint distribution $p(\eta_1, X_1) = \phi(\sqrt{n}(X_1 - \eta_1\omega))/2$.

The Gaussian integral above can be evaluated exactly and we obtain

$$\chi^2(\mathbb{E}P_0 \| \mathbb{E}P_1) = (\mathbb{E}_{X_1,\eta_1,\eta_1'} \exp(\omega^2 m \eta_1 \eta_1'))^{c_2\epsilon^{-1/s}} - 1.$$

Now, we can calculate

$$\begin{aligned}\mathbb{P}(\eta_1 = \eta_1') &= \mathbb{E}_{X_1} \frac{p(X_1|\eta_1 = 1)^2 + p(X_1|\eta_1 = -1)^2}{(p(X_1|\eta_1 = 1) + p(X_1|\eta_1 = -1))^2} \\ &= \frac{1}{2} + \frac{1}{4}\int \frac{(p(x_1|\eta_1 = 1) - p(x_1|\eta_1 = -1))^2}{p(x_1|\eta_1 = 1) + p(x_1|\eta_1 = -1)} \mathrm{d}x_1 \\ &\le \frac{1}{2} + \frac{1}{16}\sum_{b\in\{\pm 1\}} \chi^2(\mathcal{N}(b\omega, 1/n) \| \mathcal{N}(-b\omega, 1/n)) \\ &= \frac{1}{2} + \frac{\exp(4\omega^2 n) - 1}{8}.\end{aligned}$$

Together with $\mathbb{P}(\eta_1 = \eta_1') \le 1$, we have

$$\begin{aligned}\mathbb{E}_{X_1,\eta_1,\eta_1'} \exp(\omega^2 m \eta_1 \eta_1') &\le e^{-\omega^2 m} + \left(\frac{1}{2} + \frac{1}{2} \wedge \frac{e^{4\omega^2 n} - 1}{8}\right)(e^{\omega^2 m} - e^{-\omega^2 m}) \\ &= \cosh(\omega^2 m) + t\sinh(\omega^2 m),\end{aligned}$$

with $t = 1 \wedge ((e^{4\omega^2 n} - 1)/4)$. Distinguish into two scenarios:

- if $t = 1$, then $4\omega^2 n \ge 1$, and the above expression is $e^{\omega^2 m} \le e^{4\omega^4 nm}$;
- if $t < 1$, then $\omega^2 n \le 1/2$ and $t \le 8\omega^2 n$. Since $m < n$, and $\cosh(x) \le 1 + x^2, \sinh(x) \le 2x$ for all $x \in [0, 1]$, the above expression is at most

$$1 + (\omega^2 m)^2 + 2t\omega^2 m \le \exp(17\omega^4 mn).$$

Combining the above scenarios, we have

$$\chi^2(\mathbb{E}P_0 \| \mathbb{E}P_1) \le \exp(17\omega^4 nm \cdot c_2\epsilon^{-1/s}) - 1.$$

Thus, we obtain

$$1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) \gtrsim \frac{1}{1 + \chi^2(\mathbb{E}P_0 \| \mathbb{E}P_1)} \geq \exp(-17\omega^4 nm \cdot c_2 \epsilon^{-1/s}) \stackrel{!}{=} \Omega(\delta).$$

This gives the desired lower bound

$$nm \gtrsim \frac{\log(1/\delta)}{\epsilon^{\frac{4s+1}{s}}}.$$

C.4.4 Lower bounds for $\mathcal{P}_{\mathsf{D}}$

Clearly all lower bounds that apply to $\mathcal{P}_{\mathsf{Db}}$ also apply to $\mathcal{P}_{\mathsf{D}}$; in particular this gives the sample complexity lower bound for GoF. In addition, lower bounds on the minimax high-probability sample complexity of TS were derived in [62]. Hence, inspecting the claimed minimax rates, we only need to consider the problem LFHT in the cases $m \leq n \leq k$ and $n \leq m \leq k$. We give two separate constructions for the two cases, both inspired by classical constructions in the literature. As opposed to the i.i.d. sampling models, we will use the Poissonized models and rely on the formalism of pseudo-distributions as described in [62]. Specifically, suppose we can construct a random vector $(p, q) \in [0, 1]^2$ such that 1) $\mathbb{E}p = \mathbb{E}q = \Theta(1/k)$ and $|\mathbb{E}[p - q]| = \Theta(\epsilon/k)$; and 2) the following χ^2 upper bounds hold for the Poisson mixture:

$$\begin{aligned}
&\chi^2(\mathbb{E}[\mathrm{Poi}(np) \otimes \mathrm{Poi}(nq) \otimes \mathrm{Poi}(mp)] \| \mathbb{E}[\mathrm{Poi}(np) \otimes \mathrm{Poi}(nq) \otimes \mathrm{Poi}(mq)]) \leq B(n, m, \epsilon, k), \\
&\chi^2(\mathbb{E}[\mathrm{Poi}(nq) \otimes \mathrm{Poi}(np) \otimes \mathrm{Poi}(mp)] \| \mathbb{E}[\mathrm{Poi}(np) \otimes \mathrm{Poi}(nq) \otimes \mathrm{Poi}(mp)]) \leq B(n, m, \epsilon, k);
\end{aligned} \tag{C.4.4}$$

then $(n, m) \in \mathcal{R}_{\mathsf{LF}}(\epsilon, \delta, \mathcal{P}_{\mathsf{D}})$ requires $kB(n, m, \epsilon, k) \gtrsim \log(1/\delta)$ (essentially via Lemma C.4.1).

Case $m \leq n \leq k$

Suppose that $m \leq n \leq k/2$, and let p, q be two random variables defined as

$$(p, q) = \begin{cases} (\frac{1}{n}, \frac{1}{n}) & \text{with probability } \frac{n}{k}, \\ (\frac{\epsilon}{k}, \frac{2\epsilon}{k}) & \text{with probability } \frac{1}{2}(1 - \frac{n}{k}), \\ (\frac{\epsilon}{k}, 0) & \text{with probability } \frac{1}{2}(1 - \frac{n}{k}). \end{cases}$$

Note that $\mathbb{E}[p] = \mathbb{E}[q] = \Theta(1/k)$ and $|\mathbb{E}[p - q]| = \Theta(\epsilon/k)$. Let $X, Y \in \mathbb{R}^3$ be random, whose distribution is given by

$$\begin{aligned}
X|(p, q) &\sim \mathrm{Poi}(np) \otimes \mathrm{Poi}(nq) \otimes \mathrm{Poi}(mp), \\
Y|(p, q) &\sim \mathrm{Poi}(np) \otimes \mathrm{Poi}(nq) \otimes \mathrm{Poi}(mq).
\end{aligned}$$

Now, for any $(a, b, c) \in \mathbb{N}^3$ we have

$$\begin{aligned}
\mathbb{P}(X = (a, b, c)) = \frac{1}{a!b!c!} \Big(&\frac{n}{k} e^{-2-\frac{m}{n}} \left(\frac{m}{n}\right)^c + \frac{1}{2}(1 - \frac{n}{k}) e^{-(3n+m)\epsilon/k} \left(\frac{\epsilon n}{k}\right)^a \left(\frac{2\epsilon n}{k}\right)^b \left(\frac{\epsilon m}{k}\right)^c \\
&+ \frac{1}{2}(1 - \frac{n}{k}) e^{-(n+m)\epsilon/k} \left(\frac{\epsilon n}{k}\right)^a \mathbb{1}_{b=0} \left(\frac{\epsilon m}{k}\right)^c \Big).
\end{aligned}$$

Similarly, for Y we get

$$\mathbb{P}(Y=(a,b,c)) = \frac{1}{a!b!c!}\Big(\frac{n}{k}e^{-2-\frac{m}{n}}\left(\frac{m}{n}\right)^c + \frac{1}{2}(1-\frac{n}{k})e^{-(3n+2m)\epsilon/k}\left(\frac{\epsilon n}{k}\right)^a\left(\frac{2\epsilon n}{k}\right)^b\left(\frac{2\epsilon m}{k}\right)^c + \frac{1}{2}(1-\frac{n}{k})e^{-n\epsilon/k}\left(\frac{\epsilon n}{k}\right)^a \mathbb{1}_{b=c=0}\Big).$$

In particular, we have

$$\mathbb{P}(Y=(a,b,c)) = \Omega\left(\frac{1}{a!b!c!}\right)\begin{cases}1 & \text{if } (a,b,c)=(0,0,0),\\ \frac{n}{k}\left(\frac{m}{n}\right)^c & \text{otherwise.}\end{cases}$$

Notice also that

$$\begin{aligned}&\mathbb{P}(X=(a,b,c)) - \mathbb{P}(Y=(a,b,c))\\ &= \frac{1}{a!b!c!}\underbrace{\frac{1}{2}(1-\frac{n}{k})e^{-n\epsilon/k}}_{=\Theta(1)}\left(\frac{\epsilon}{k}\right)^{a+b+c} n^{a+b}m^c\\ &\qquad\times\underbrace{\left[2^b e^{-(2n+m)\epsilon/k}\left(1-2^c e^{-m\epsilon/k}\right)+\mathbb{1}_{b=0}\left(e^{-m\epsilon/k}-\mathbb{1}_{c=0}\right)\right]}_{=:I_{bc}}\\ &= \frac{\Theta(1)}{a!b!c!}\left(\frac{\epsilon}{k}\right)^{a+b+c} n^{a+b}m^c I_{bc},\end{aligned}$$

where

$$|I_{bc}| \lesssim \begin{cases}\frac{nm\epsilon^2}{k^2} & \text{if } b=c=0,\\ \frac{2^b m\epsilon}{k} & \text{if } b\geq 1, c=0,\\ \frac{n\epsilon}{k} & \text{if } b=0, c=1,\\ 2^{b+c} & \text{otherwise.}\end{cases} \tag{C.4.5}$$

We now turn to bounding the χ^2-divergence between X and Y. Using the estimates (C.4.5), we obtain

$$\begin{aligned}\chi^2(X\|Y) &= \sum_{(a,b,c)\in\mathbb{N}^3}\frac{(\mathbb{P}(X=(a,b,c))-\mathbb{P}(Y=(a,b,c)))^2}{\mathbb{P}(Y=(a,b,c))}\\ &\lesssim I_{00}^2 + \left(\sum_{b=c=0,a\geq 1}+\sum_{a\geq 0,b+c\geq 1}\right)\frac{\frac{1}{a!b!c!}\left(\frac{\epsilon}{k}\right)^{2a+2b+2c}n^{2a+2b}m^{2c}I_{bc}^2}{\frac{n}{k}\left(\frac{m}{n}\right)^c}\\ &= I_{00}^2\left(1+\sum_{a\geq 1}\frac{1}{a!}\frac{\epsilon^{2a}n^{2a-1}}{k^{2a-1}}\right)+\left(\sum_{a\geq 0}\frac{1}{a!}\frac{\epsilon^{2a}n^{2a}}{k^{2a}}\right)\sum_{b+c\geq 1}\frac{1}{b!c!}\frac{\epsilon^{2b+2c}n^{2b+c-1}m^c}{k^{2b+2c-1}}I_{bc}^2\\ &\lesssim \frac{n^2m^2\epsilon^4}{k^4}\underbrace{\left(1+\frac{n\epsilon^2}{k}e^{\epsilon^2n^2/k^2}\right)}_{=\Theta(1)}+\underbrace{e^{\epsilon^2n^2/k^2}}_{=\Theta(1)}\sum_{b+c\geq 1}\frac{1}{b!c!}\frac{\epsilon^{2b+2c}n^{2b+c-1}m^c}{k^{2b+2c-1}}I_{bc}^2.\end{aligned}$$

Focusing on the sum and decomposing it as $\sum_{b+c\geq 1} = \sum_{c=0,b\geq 1} + \sum_{b=0,c=1} + \sum_{b=0,c\geq 2} + \sum_{b,c\geq 1}$ we have the estimates

$$\begin{aligned}
&\sum_{b+c\geq 1} \frac{1}{b!c!} \frac{\epsilon^{2b+2c} n^{2b+c-1} m^c}{k^{2b+2c-1}} I_{bc}^2 \\
&\lesssim \sum_{c=0,b\geq 1} \frac{1}{b!} \frac{\epsilon^{2b+2} n^{2b-1} 4^b m^2}{k^{2b+1}} + \frac{\epsilon^4 m n^2}{k^3} + \sum_{b=0,c\geq 2} \frac{1}{c!} \frac{\epsilon^{2c} n^{c-1} m^c 4^c}{k^{2c-1}} + \sum_{b,c\geq 1} \frac{1}{b!c!} \frac{\epsilon^{2b+2c} n^{2b+c-1} m^c 4^{b+c}}{k^{2b+2c-1}} \\
&\lesssim \frac{\epsilon^4 m^2 n}{k^3} + \frac{\epsilon^4 m n^2}{k^3} + \frac{\epsilon^4 m^2 n}{k^3} + \frac{\epsilon^4 m n^2}{k^3} \lesssim \frac{\epsilon^4 m n^2}{k^3}.
\end{aligned}$$

As $m \leq k$, we obtain

$$\chi^2(X\|Y) \lesssim \frac{\epsilon^4 m n^2}{k^3}.$$

By (C.4.4) we conclude that in the regime $m \leq n \leq k$, $(n,m) \in \mathcal{R}_{\mathsf{LF}}(\epsilon, \delta, \mathcal{P}_{\mathsf{D}})$ requires $n^2 m \gtrsim k^2 \log(1/\delta)/\epsilon^4$, as desired.

Case $n \leq m \leq k$

This case is entirely analogous to the previous case with minor modifications. Suppose that $n \leq m \leq k/2$, and let p, q be two random variables defined as

$$(p,q) = \begin{cases} (\frac{1}{m}, \frac{1}{m}) & \text{with probability } \frac{m}{k}, \\ (\frac{\epsilon}{k}, \frac{2\epsilon}{k}) & \text{with probability } \frac{1}{2}(1 - \frac{m}{k}), \\ (\frac{\epsilon}{k}, 0) & \text{with probability } \frac{1}{2}(1 - \frac{m}{k}). \end{cases}$$

Let $X, Y \in \mathbb{R}^3$ be random, whose distribution is given by

$$\begin{aligned}
X|(p,q) &\sim \mathrm{Poi}(np) \otimes \mathrm{Poi}(nq) \otimes \mathrm{Poi}(mp), \\
Y|(p,q) &\sim \mathrm{Poi}(nq) \otimes \mathrm{Poi}(np) \otimes \mathrm{Poi}(mp).
\end{aligned}$$

Now, for any $(a,b,c) \in \mathbb{N}^3$ we have

$$\begin{aligned}
\mathbb{P}(X = (a,b,c)) = \frac{1}{a!b!c!} \Big(&\frac{m}{k} e^{-\frac{2n}{m} - 1} \left(\frac{n}{m}\right)^{a+b} + \frac{1}{2}(1 - \frac{m}{k}) e^{-(3n+m)\epsilon/k} \left(\frac{\epsilon n}{k}\right)^a \left(\frac{2\epsilon n}{k}\right)^b \left(\frac{\epsilon m}{k}\right)^c \\
&+ \frac{1}{2}(1 - \frac{m}{k}) e^{-(n+m)\epsilon/k} \left(\frac{\epsilon n}{k}\right)^a \mathbb{1}_{b=0} \left(\frac{\epsilon m}{k}\right)^c \Big).
\end{aligned}$$

Similarly, for Y we get

$$\begin{aligned}
\mathbb{P}(Y = (a,b,c)) = \frac{1}{a!b!c!} \Big(&\frac{m}{k} e^{-\frac{2n}{m} - 1} \left(\frac{n}{m}\right)^{a+b} + \frac{1}{2}(1 - \frac{m}{k}) e^{-(3n+m)\epsilon/k} \left(\frac{2\epsilon n}{k}\right)^a \left(\frac{\epsilon n}{k}\right)^b \left(\frac{\epsilon m}{k}\right)^c \\
&+ \frac{1}{2}(1 - \frac{m}{k}) e^{-(n+m)\epsilon/k} \mathbb{1}_{a=0} \left(\frac{\epsilon n}{k}\right)^b \left(\frac{\epsilon m}{k}\right)^c \Big).
\end{aligned}$$

In particular, we have

$$\mathbb{P}(Y = (a,b,c)) = \Omega\left(\frac{1}{a!b!c!}\right) \begin{cases} 1 & \text{if } (a,b,c) = (0,0,0), \\ \frac{m}{k} \left(\frac{n}{m}\right)^{a+b} & \text{otherwise.} \end{cases}$$

Notice that

$$\begin{aligned}
&\mathbb{P}(X=(a,b,c))-\mathbb{P}(Y=(a,b,c))\\
&\quad=\frac{1}{a!b!c!}\frac{1}{2}(1-\frac{m}{k})e^{-(n+m)\epsilon/k}\left(\frac{\epsilon}{k}\right)^{a+b+c}n^{a+b}m^{c}\underbrace{\left(e^{-2n\epsilon/k}(2^b-2^a)+\mathbb{1}_{b=0}-\mathbb{1}_{a=0}\right)}_{=:J_{ab}}\\
&\quad=\frac{\Theta(1)}{a!b!c!}\left(\frac{\epsilon}{k}\right)^{a+b+c}n^{a+b}m^{c}J_{ab},
\end{aligned}$$

where

$$|J_{ab}|\lesssim\begin{cases}0 & \text{if } a+b=0,\\ \frac{n\epsilon}{k} & \text{if } a+b=1,\\ 2^{a+b} & \text{if } a+b\geq 2.\end{cases}\tag{C.4.6}$$

We now turn to bounding the χ^2-divergence between X and Y. We have

$$\begin{aligned}
\chi^2(X\|Y)&=\sum_{(a,b,c)\in\mathbb{N}^3}\frac{\left(\mathbb{P}(X=(a,b,c))-\mathbb{P}(Y=(a,b,c))\right)^2}{\mathbb{P}(Y=(a,b,c))}\\
&\asymp\sum_{a+b+c\geq 1}\frac{\frac{1}{a!^2b!^2c!^2}\left(\frac{\epsilon}{k}\right)^{2a+2b+2c}n^{2a+2b}m^{2c}J_{ab}^2}{\frac{1}{a!b!c!}\frac{m}{k}\left(\frac{n}{m}\right)^{a+b}}\\
&\asymp\sum_{a+b+c\geq 1}\frac{1}{a!b!c!}\frac{\epsilon^{2a+2b+2c}n^{a+b}m^{2c+a+b-1}J_{ab}^2}{k^{2a+2b+2c-1}}\\
&=\underbrace{e^{\epsilon^2m^2/k^2}}_{\Theta(1)}\sum_{a+b\geq 1}\frac{1}{a!b!}\frac{\epsilon^{2a+2b}n^{a+b}m^{a+b-1}J_{ab}^2}{k^{2a+2b-1}},
\end{aligned}$$

where the last step follows from $J_{ab}=0$ if $a=b=0$. Now writing $t=a+b$ and distinguishing into cases $t=1$ and $t\geq 2$, by (C.4.6) we have

$$\chi^2(X\|Y)\lesssim\frac{\epsilon^4n^3}{k^3}+\sum_{t\geq 2}\frac{2^t}{t!}\frac{\epsilon^{2t}n^tm^{t-1}4^t}{k^{2t-1}}\lesssim\frac{\epsilon^4n^3}{k^3}+\frac{\epsilon^4n^2m}{k^3}\lesssim\frac{\epsilon^4n^2m}{k^3},$$

where the last line uses that $n\leq m$. Once again, we can conclude by (C.4.4) that $n^2m\gtrsim\log(1/\delta)k^2/\epsilon^4$ is a lower bound for the sample complexity of LFHT.

Appendix D

Appendix of "Density Estimation Using the Perceptron"

D.1 Auxiliary Technical Results

In this section we list some technical lemmas that are used in our later proofs.

Theorem D.1.1 (Plancherel theorem)**.** *Let $f, g \in L^2(\mathbb{R}^d)$. Then*

$$\int_{\mathbb{R}^d} f(x)\overline{g(x)}\mathrm{d}x = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \widehat{f}(\omega)\overline{\widehat{g}(\omega)}\mathrm{d}\omega.$$

Lemma D.1.2. *Suppose $t, x, y > 0$. Then there exist finite t-dependent constants C_1, C_2 such that*

$$x \leq\ y\log(3+1/y)^t \implies \frac{x}{\log(3+1/x)^t} \leq C_1\ y \implies x \leq C_2\ y\log(3+1/y)^t.$$

Proof. Let us focus on the first implication. If $x \leq y$, then it clearly holds. If $y \leq x \leq y\log(3+1/y)^t$ then it suffices to show

$$\left(\frac{x}{\log(3+1/x)^t} \leq\right) \frac{y\log(3+1/y)^t}{\log(3+1/(y\log(3+1/y)^t))^t} \overset{!}{\leq} C_1 y.$$

The inequality marked by ! is equivalent to

$$3+1/y \leq (3+1/(y\log(3+1/y)^t))^{\sqrt[t]{C_1}}.$$

Now, if $y \geq 1/2$ then clearly taking $C_1 = \log_3(5)^t$ works. Suppose that instead $y \in (0, 1/2)$. Then, since log grows slower than any polynomial, there exists a t-dependent constant $c_t < \infty$ such that $\log(3+1/y) \leq c_t y^{-1/(2t)}$ for all $y \in (0, 1/2)$. Therefore, we have

$$3 + \frac{1}{y\log(3+1/y)^t} \geq 3 + \frac{1}{c_t^t y^{1/2}}.$$

It is then clear that

$$3 + \frac{1}{y} \leq \left(3 + \frac{1}{c_t^t y^{1/2}}\right)^{\sqrt[t]{C_1}}$$

holds for all $y \in (0, 1/2)$ if we take C_1 large enough in terms of t. The second implication follows analogously and we omit its proof. □

Lemma D.1.3. *Let μ be a probability distribution on $\mathbb{R}^d$ and $\gamma \in (0, 2)$. Then*

$$\mathbb{E}_{X\sim\mu}\int_{\mathbb{R}^d} \frac{(\cos\langle\omega, X\rangle - 1)^2 + \sin^2\langle\omega, X\rangle}{\|\omega\|^{d+\gamma}}\mathrm{d}\omega \leq \frac{16\pi^{d/2}M_\gamma(\mu)}{\Gamma(d/2)\gamma(2-\gamma)}.$$

Proof. We use the inequalities $(\cos t - 1)^2 + \sin^2(t) \leq 4(t^2 \wedge 1)$ valid for all $t \in \mathbb{R}$. Plugging in and using the Cauchy-Schwarz inequality, the quantity on the left hand side can be bounded as

$$\begin{aligned}
4\mathbb{E}\int_{\mathbb{R}^d} \frac{1 \wedge (\|\omega\|^2\|X\|^2)}{\|\omega\|^{d+\gamma}}\mathrm{d}\omega &\leq 4\,\mathrm{vol}_{d-1}(\mathbb{S}^{d-1})\mathbb{E}\int_0^\infty \frac{1 \wedge (r^2\|X\|^2)}{r^{1+\gamma}}\mathrm{d}r \\
&= \frac{8\pi^{d/2}}{\Gamma(\frac{d}{2})}\mathbb{E}\left\{\|X\|^2\int_0^{\|X\|^{-1}} \frac{1}{r^{\gamma-1}}\mathrm{d}r + \int_{\|X\|^{-1}}^\infty \frac{1}{r^{1+\gamma}}\mathrm{d}r\right\} \\
&= \frac{16\pi^{d/2}M_\gamma(\nu)}{\Gamma(\frac{d}{2})\gamma(2-\gamma)},
\end{aligned}$$

where $\mathrm{vol}_{d-1}(\mathbb{S}^{d-1}) = \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)}$ is the surface area of the unit $(d-1)$-sphere. □

Lemma D.1.4. *For $\gamma \in (0,2)$ define*

$$B_\gamma = \begin{cases} \sup_{0<a<c} \left| \int_a^c \frac{\sin(\omega)}{\omega} \mathrm{d}\omega \right| & \text{if } \gamma = 1, \\ \sup_{0<a<c} \left| \int_a^c \frac{\cos(\omega)}{\omega^{(1+\gamma)/2}} \mathrm{d}\omega \right| & \text{if } \gamma \in (0,1), \\ \sup_{0<a<c} \left| \int_a^c \frac{\cos(\omega)-1}{\omega^{(1+\gamma)/2}} \mathrm{d}\omega \right| & \text{if } \gamma \in (1,2). \end{cases}$$

Then $B_\gamma < \infty$.

Proof. In the $\gamma > 1$ case, one immediately has $B_\gamma \leq \int_0^\infty \left| \frac{\min\{1,\omega^2/2\}}{\omega^{(1+\gamma)/2}} \right| \mathrm{d}\omega < \infty$. Now let us consider $\gamma \leq 1$. One has that $B_\gamma = \sup_{0<c_1<c_2} |I_\gamma(c_2) - I_\gamma(c_1)| \leq \sup_{c>0} 2|I_\gamma(c)|$ where

$$I_\gamma(a) = \begin{cases} \int_1^a \frac{\sin(\omega)}{\omega} \mathrm{d}\omega & \text{if } \gamma = 1, \\ \int_1^a \frac{\cos(\omega)}{\omega^{(1+\gamma)/2}} \mathrm{d}\omega & \text{if } \gamma \in (0,1), \end{cases}$$

Clearly $a \mapsto I_\gamma(a)$ is continuous on $(0,\infty)$ for each $\gamma \in (0,2)$. Moreover,

$$|I_\gamma(a)| \leq |a-1| \cdot \max\{a^{-(\gamma+1)/2}, 1\}.$$

Therefore, we only need to show that $\lim_{a\to\infty} |I_\gamma(a)|$ and $\lim_{a\to 0} |I_\gamma(a)|$ are both finite. Let $g_\gamma(x) = x^{-(\gamma+1)/2}$ and $f_\gamma(x) = \sin(x)$ if $\gamma = 1$ and $\cos(x)$ otherwise, so that we may write $I_\gamma(a) = \int_a^1 f_\gamma(\omega) g_\gamma(\omega) \mathrm{d}\omega$.

1. For $a \to \infty$, since $g_\gamma(\infty) = 0$ and $|\int_1^a f_\gamma(x)\mathrm{d}t| \leq 2$ is uniformly bounded, $\int_1^\infty f(\omega)g(\omega)\mathrm{d}\omega$ converges to a finite value by Dirichlet's test for improper integrals [145, page 391].[1]
2. For $a \to 0$, the conclusion follows by the inequality $|\sin(\omega)| \leq \min\{|\omega|, 1\}$ in the case $\gamma = 1$, and by $|I_\gamma(a)| \leq \int_a^1 \omega^{-(1+\gamma)/2}\mathrm{d}\omega \leq \int_0^1 \omega^{-(1+\gamma)/2}\mathrm{d}\omega = \frac{2}{1-\gamma}$ for $\gamma < 1$.

Therefore, $\sup_{a>0} |I_\gamma(a)| < \infty$ which concludes the proof.

□

Lemma D.1.5. *Let $\int_0^\infty \cdot \, \mathrm{d}\omega =: \lim_{\epsilon\to 0} \int_{1/\epsilon \geq \omega \geq \epsilon} \cdot \, \mathrm{d}\omega$ and recall the definition of ψ_γ from (5.2.6). Then, for $x \neq 0$ the following hold:*

$$\psi_\gamma(x) = C_{\psi_\gamma} \begin{cases} \int_0^\infty \frac{\sin(\omega x)}{\omega} \mathrm{d}\omega + \frac{\pi}{2} & \text{for } \gamma = 1, \\ \int_0^\infty \frac{\cos(\omega x)}{\omega^{(1+\gamma)/2}} \mathrm{d}\omega & \text{for } \gamma \in (0,1), \\ \int_0^\infty \frac{\cos(\omega x)-1}{\omega^{(1+\gamma)/2}} \mathrm{d}\omega & \text{for } \gamma \in (1,2), \end{cases}$$

where

$$C_{\psi_\gamma} = \begin{cases} \left(\cos\left(\frac{\pi(\gamma-1)}{4}\right) \Gamma\left(\frac{1-\gamma}{2}\right) \right)^{-1} & \text{if } \gamma \neq 1, \\ \frac{1}{\pi} & \text{if } \gamma = 1. \end{cases}$$

[1] Reference pointed out by user Siminore on math.stackexchange.com.

Proof. For $x \neq 0$ clearly

$$\int_0^\infty \frac{\sin(\omega x)}{w}\mathrm{d}w = \mathrm{sign}(x)\int_0^\infty \frac{\sin(\omega)}{w}\mathrm{d}w = \mathrm{sign}(x)\frac{\pi}{2},$$

which shows the first claim. Assume from here on without loss of generality that $x > 0$. For $\gamma \in (0,1)$, by the residue theorem,

$$\begin{aligned}
\int_0^\infty \frac{\cos(\omega x)}{\omega^{(1+\gamma)/2}}\mathrm{d}\omega &= x^{(\gamma-1)/2}\int_0^\infty \Re\left(\frac{e^{i\omega}}{\omega^{(1+\gamma)/2}}\right)\mathrm{d}w \\
&= x^{(\gamma-1)/2}\,\Re\left(ie^{-i\frac{\pi}{2}\gamma}\right)\int_0^\infty \frac{e^{-z}}{z^{(1+\gamma)/2}}\mathrm{d}z \\
&= x^{(\gamma-1)/2}\cos\left(\frac{\pi(\gamma-1)}{4}\right)\Gamma((1-\gamma)/2).
\end{aligned}$$

Similarly, for $\gamma \in (1,2)$, integration by parts and the residue theorem gives

$$\begin{aligned}
\int_0^\infty \frac{\cos(\omega x)-1}{\omega^{(1+\gamma)/2}}\mathrm{d}\omega &= x^{(\gamma-1)/2}\int_0^\infty (\cos(\omega)-1)\mathrm{d}\left(\frac{-1}{((\gamma-1)/2)\,\omega^{(\gamma-1)/2}}\right) \\
&= -x^{(\gamma-1)/2}\int_0^\infty \frac{\sin(w)}{(\gamma-1)/2\,\omega^{(\gamma-1)/2}}\mathrm{d}\omega \\
&= -x^{(\gamma-1)/2}\frac{2}{\gamma-1}\int_0^\infty \mathsf{im}\left(\frac{e^{iw}}{\omega^{(\gamma-1)/2}}\right)\mathrm{d}\omega \\
&= -x^{(\gamma-1)/2}\frac{2}{\gamma-1}\,\mathsf{im}\left(ie^{-\frac{\pi}{2}(\gamma-1)/2}\right)\int_0^\infty \frac{e^{-z}}{z^{(\gamma-1)/2}}\mathrm{d}z \\
&= -x^{(\gamma-1)/2}\frac{2}{\gamma-1}\cos\left(\frac{\pi(\gamma-1)}{4}\right)\Gamma(1-(\gamma-1)/2) \\
&= x^{(\gamma-1)/2}\cos\left(\frac{\pi(\gamma-1)}{4}\right)\Gamma((1-\gamma)/2).
\end{aligned}$$

□

Lemma D.1.6. *Let ϕ be the probability density function of $\mathcal{N}(0,\sigma I_d)$ and write $\widehat{\phi}$ for its Fourier transform. Then, for any $\beta \geq 1$,*

$$\|\widehat{\phi}(\omega)\|\omega\|^\beta\|_2^2 = \frac{\pi^{d/2}}{\Gamma(d/2)\sigma^{2\beta+d}}\Gamma\left(\frac{2\beta+d}{2}\right) \leq \frac{5\pi^{d/2}}{\Gamma(d/2)\sigma^{2\beta+d}}\left(\frac{2\beta+d}{2e}\right)^{\frac{2\beta+d-1}{2}}.$$

Proof. It is well known that $\widehat{\phi}(\omega) = e^{-\frac{\sigma^2}{2}\|\omega\|^2}$. The claimed equality then follows from the formula for the 2β'th moment of the Gaussian distribution with mean 0 and variance $1/(2\sigma^2)$. The inequality follows by Lemma D.1.7.

□

Lemma D.1.7 (Properties of the gamma function)**.** *For all $x > 1$ the inequality $\Gamma(x) \leq 5(x/e)^{x-1/2}$ holds.*

Proof. In [148] authors showed that $\log\Gamma(x) \leq (x-\frac{1}{2})\log(x) - x + \frac{1}{2}\log(2\pi) + 1$ for all $x \geq 1$, from which the second claim follows as $\exp(\frac{1}{2}\log(2\pi) + 1/2) < 5$. □

Lemma D.1.8. *Let $b \geq 1$ and $|a| < b$. Then*

$$\int_0^r x^a |\sin(x)|^b dx = O(r^{a+b})$$

as $r \to \infty$, where we hide constants depending on a, b.

Proof. Since we are only interested in the asymptotic behaviour as $r \to \infty$, assume without loss of generality that $r \geq 1$. Then, we have

$$\int_0^r x^a |\sin(x)|^b \mathrm{d}x = \underbrace{\int_0^1 x^a |\sin(x)|^b \mathrm{d}x}_{I} + \underbrace{\int_1^r x^a |\sin(x)|^b \mathrm{d}x}_{II}.$$

Using the inequality $|\sin(x)| \leq x$, we can bound the first term as

$$I \leq \int_0^1 x^{a+b} \mathrm{d}x = \frac{1}{a+b+1} \leq 1 = O(r^{a+b}),$$

since $a + b > 0$. For the second term, we obtain

$$II \leq \int_1^r x^a \mathrm{d}x = \left\{ \begin{array}{ll} \frac{r^{a+1}-1}{a+1} & \text{if } a \neq -1 \\ \log(r) & \text{if } a = -1 \end{array} \right\} \quad = O(r^{a+b}),$$

where the last step uses $a + b > 0$ and $b \geq 1$.

□

Lemma D.1.9. *Let $a, b, c \in \mathbb{R}$ with $b > 0$ be constants. For all large enough r one has*

$$\int_r^\infty x^a \exp\left(-\frac{bx}{\log^2(x+2)}\right) dx < r^{-c}.$$

Proof. Assume, without loss of generality, that $c \geq 0$. For all large enough x one has

$$\exp\left(-\frac{bx}{\log^2(x+2)}\right) < x^{-a-c-2}.$$

Therefore, for large enough r,

$$\int_r^\infty x^a \exp\left(-\frac{bx}{\log^2(x+2)}\right) \mathrm{d}x < \int_r^\infty x^{-c-2} \mathrm{d}x \asymp r^{-c-1} < r^{-c}.$$

□

Lemma D.1.10. *Let J_ν be the Bessel function of the first kind of order ν.*

1. *For all* $x \in \mathbb{R}^d$,
$$\int_{\mathbb{R}^d} e^{i\langle x,v\rangle} \mathrm{d}\sigma(v) = (2\pi)^{d/2} \|x\|^{1-d/2} J_{d/2-1}(\|x\|).$$

2. *For any* $\nu \in \mathbb{R}$, *as* $x \to \infty$
$$J_\nu(x) = \sqrt{\frac{2}{\pi x}} \cos\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4}\right) + O(x^{-3/2}). \tag{D.1.1}$$

3. *For all* $x \in \mathbb{R}^d$,
$$\int_{\mathbb{B}^d(0,1)} e^{i\langle x,w\rangle}\, dw = \left(\frac{2\pi}{\|x\|}\right)^{d/2} J_{d/2}(\|x\|).$$

Proof. For Item 1 set $r = \|x\|$ and $s = (2\pi)^{-1}$ in the calculation on page 154 of [186].

For Item 2 see [203, Eq. (1) in Section 7.21].

For Item 3, we can compute
$$\begin{aligned}\int_{\|w\|\le 1} e^{i\langle x,w\rangle} \mathrm{d}w &= \int_{-1}^{1} e^{i\|x\|w_1} \int_{w_2^2+\cdots+w_d^2 \le 1-w_1^2} \mathrm{d}w_2 \ldots \mathrm{d}w_d \mathrm{d}w_1 \\ &= \frac{\pi^{(d-1)/2}}{\Gamma(\frac{d+1}{2})} \int_{-1}^{1} e^{i\|x\|w_1} (1-w_1^2)^{(d-1)/2} \mathrm{d}w_1. \end{aligned} \tag{D.1.2}$$

Recall from [203, Section 3.1] the definition of the Bessel function of the first kind as
$$J_\nu(x) = \frac{(x/2)^\nu}{\Gamma(\nu+\frac{1}{2})\Gamma(\frac{1}{2})} \int_0^\infty \cos(x\cos(\theta)) \sin^{2\nu}(\theta) \mathrm{d}\theta$$
valid for $\nu > -1/2$; the above is also known as the Poisson representation. Changing variables to $u = \cos(\theta)$, we see that it is equal to
$$J_\nu(x) = \frac{(x/2)^\nu}{\Gamma(\nu+\frac{1}{2})\Gamma(\frac{1}{2})} \int_{-1}^{1} e^{ixu} (1-u^2)^{\nu-1/2} \mathrm{d}u. \tag{D.1.3}$$
Comparing (D.1.3) with (D.1.2) concludes the proof. □

Lemma D.1.11. *There exists a radial function* $h_0 \in L^2(\mathbb{R}^d)$ *such that*
$$\begin{aligned} &\mathrm{supp}(h_0) \subseteq \mathbb{B}(0,1), \\ &|\widehat{h}_0(w)| \le C \exp\left(-\frac{c\|w\|}{\log(\|w\|+2)^2}\right) && \text{for all } w \in \mathbb{R}^d, \\ &|\widehat{h}_0(w)| \ge \frac{1}{2} && \text{for all } \|w\| \le r_{\min}, \end{aligned}$$
where $C, c, r_{\min} > 0$.

Proof. Apply Theorem 1.4 in [47] using the spherically symmetric weight function $u : \mathbb{R}^d \to \mathbb{R}_{\le 0}$ defined by
$$u(w) = u(\|w\|) = -\frac{\|w\|}{\log(\|w\|+2)^2} \left(\frac{(\|w\|-2)_+}{\|w\|+2}\right)^4,$$
where $(a)_+ := \max(a, 0)$ for $a \in \mathbb{R}$. □

D.2 Proof of Proposition 5.2.5

For $v \in \mathbb{S}^{d-1}$ and $b \in \mathbb{R}$ let $\theta_v(x) = \langle v, x\rangle$ and write $\eta_v =: \theta_v \#(\mu - \nu)$ for the pushforward of the measure $\mu - \nu$ through the map θ_v. To start with, we notice that

$$\begin{aligned} &\int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \Big[\mathbb{E}\psi_\gamma(\langle X, v\rangle - b) - \mathbb{E}\psi_\gamma(\langle Y, v\rangle - b)\Big]^2 \mathrm{d}b\mathrm{d}\sigma(v) \\ &= \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \psi_\gamma(x-b)\mathrm{d}\eta_v(x)\right)^2 \mathrm{d}b\mathrm{d}\sigma(v), \end{aligned} \tag{D.2.1}$$

For each $v \in \mathbb{S}^{d-1}$, the measure η_v has at most countably many atoms, therefore $b \mapsto \eta_v(\{b\}) = 0$ Leb-almost everywhere. Then, by Tonelli's theorem we can conclude that $\eta_v(\{b\}) = 0$ for $\sigma \otimes$ Leb-almost every (v, b), thus going forward we can focus on the case $x \neq b$. By Lemma D.1.5, and writing $A_\epsilon = [\epsilon, 1/\epsilon]$ for $\epsilon > 0$, we have

$$\int_{\mathbb{R}} \psi_\gamma(x-b)\mathrm{d}\eta_v(x) = C_{\psi_\gamma} \int_{\mathbb{R}} \lim_{\epsilon\to 0} \int_{A_\epsilon} \left\{ \begin{array}{ll} \dfrac{\sin(\omega(x-b))}{\omega} & \text{if } \gamma = 1 \\ \dfrac{\cos(\omega(x-b))}{\omega^{(1+\gamma)/2}} & \text{if } \gamma \in (0,1) \\ \dfrac{\cos(\omega(x-b)) - 1}{\omega^{(1+\gamma)/2}} & \text{if } \gamma \in (1,2) \end{array} \right\} \mathrm{d}\omega\mathrm{d}\eta_v(x).$$

Note that in the $\gamma = 1$ case we implicitly used that $\int \mathrm{d}\eta_v(x) = 0$. To exchange the integral over x and the limit over ϵ, notice that for any $\epsilon > 0$ and $x \neq b \in \mathbb{R}$,

$$\begin{aligned} &\left|\int_\epsilon^{1/\epsilon} \frac{\sin(\omega(x-b))}{\omega}\mathrm{d}\omega\right| \leq B_\gamma && \text{if } \gamma = 1, \\ &\left|\int_\epsilon^{1/\epsilon} \frac{\cos(\omega(x-b))}{\omega^{(1+\gamma)/2}}\mathrm{d}\omega\right| \leq B_\gamma |x-b|^{(\gamma-1)/2} && \text{if } \gamma \in (0,1), \\ &\left|\int_\epsilon^{1/\epsilon} \frac{\cos(\omega(x-b)) - 1}{\omega^{(1+\gamma)/2}}\mathrm{d}\omega\right| \leq B_\gamma |x-b|^{(\gamma-1)/2} && \text{if } \gamma \in (1,2). \end{aligned}$$

where $B_\gamma < \infty$ depends only on γ and is defined in Lemma D.1.4. We now show that $\int_{\mathbb{R}} |x-b|^{(\gamma-1)/2}\mathrm{d}|\eta_v|(x) < \infty$ for $\sigma \otimes$ Leb-almost every b, v. To this end, let $S = \{(b, v) \in \mathbb{R} \times \mathbb{S}^{d-1} : \int_{\mathbb{R}} |x-b|^{(\gamma-1)/2}\mathrm{d}|\eta_v|(x) = \infty\}$ and assume for contradiction $(\sigma \otimes \mathrm{Leb})(S) > 0$. Then $\mathbb{1}_{([-B,B]\times\mathbb{S}^{d-1})\cap S} \uparrow \mathbb{1}_S$ as $B \to \infty$, and thus by the monotone convergence theorem there exists a finite B such that $\mathrm{Leb}(([-B, B] \times \mathbb{S}^{d-1}) \cap S) > 0$. However, by Tonelli's theorem we

have

$$\begin{aligned}\int_{-B}^{B}\left(\int_{\mathbb{R}}|x-b|^{(\gamma-1)/2}\mathrm{d}|\eta_v|(x)\right)^2\mathrm{d}b &\le \int_{-B}^{B}\int_{\mathbb{R}}|x-b|^{\gamma-1}\mathrm{d}|\eta_v|(x)\mathrm{d}b\\ &\le 2\int_{\mathbb{R}}\int_0^{B+|x|}b^{\gamma-1}\mathrm{d}b\mathrm{d}|\eta_v|(x)\\ &\lesssim \int_{\mathbb{R}}(B+|x|)^{\gamma}\mathrm{d}|\eta_v|(x)\\ &\lesssim B^{\gamma}+\mathbb{E}_{X\sim\mu}\left[|\langle v,X\rangle|^{\gamma}\right]+\mathbb{E}_{Y\sim\nu}\left[|\langle v,Y\rangle|^{\gamma}\right]\\ &\le B^{\gamma}+M_{\gamma}(\mu+\nu),\end{aligned}$$

which, after integration over $v\in\mathbb{S}^{d-1}$, leads to a contradiction if $M_\gamma(\mu+\nu)<\infty$. Continuing under the assumption $M_\gamma(\mu+\nu)<\infty$, we can apply the dominated convergence theorem to obtain

$$\int_{\mathbb{R}}\psi_\gamma(x-b)\mathrm{d}\eta_v(x)=C_{\psi_\gamma}\lim_{\epsilon\to0}\int_{\mathbb{R}}\int_{A_\epsilon}\left\{\begin{array}{ll}\dfrac{\sin(\omega(x-b))}{\omega} & \text{if } \gamma=1\\ \dfrac{\cos(\omega(x-b))}{\omega^{(1+\gamma)/2}} & \text{if } \gamma\in(0,1)\\ \dfrac{\cos(\omega(x-b))-1}{\omega^{(1+\gamma)/2}} & \text{if } \gamma\in(1,2)\end{array}\right\}\mathrm{d}\omega\mathrm{d}\eta_v(x).$$

Then by Fubini's theorem, we exchange the order of integration to get

$$\int_{\mathbb{R}}\psi_\gamma(x-b)\mathrm{d}\eta_v(x)=C_{\psi_\gamma}\lim_{\epsilon\to0}\int_{A_\epsilon}\int_{\mathbb{R}}\left\{\begin{array}{ll}\dfrac{\sin(\omega(x-b))}{\omega} & \text{if } \gamma=1\\ \dfrac{\cos(\omega(x-b))}{\omega^{(1+\gamma)/2}} & \text{if } \gamma\in(0,1)\\ \dfrac{\cos(\omega(x-b))-1}{\omega^{(1+\gamma)/2}} & \text{if } \gamma\in(1,2)\end{array}\right\}\mathrm{d}\eta_v(x)\mathrm{d}\omega.$$

Notice that $\int_{\mathbb{R}}e^{-i\omega x}\mathrm{d}\eta_v(x)=\widehat{\eta}_v(\omega)$, $\widehat{\eta}_v(\omega)=\overline{\widehat{\eta}_v(-\omega)}$ and $\widehat{\eta}_v(0)=0$,

$$\begin{aligned}\int_{\mathbb{R}}\psi_\gamma(x-b)\mathrm{d}\eta_v(x)&=C_{\psi_\gamma}\lim_{\epsilon\to0}\int_{A_\epsilon}\frac{1}{\omega^{(1+\gamma)/2}}\left\{\begin{array}{ll}\mathsf{im}(e^{-i\omega b}\overline{\widehat{\eta}_v(\omega)}) & \text{if } \gamma=1\\ \Re(e^{-i\omega b}\overline{\widehat{\eta}_v(\omega)}) & \text{if } \gamma\neq1\end{array}\right\}\mathrm{d}\omega\\ &=C_{\psi_\gamma}\lim_{\epsilon\to0}\left\{\begin{array}{ll}\mathsf{im}\left(\widehat{\Psi}_{\gamma,v,\epsilon}(b)\right) & \text{if } \gamma=1\\ \Re\left(\widehat{\Psi}_{\gamma,v,\epsilon}(b)\right) & \text{if } \gamma\neq1.\end{array}\right.\end{aligned}$$

where we write

$$\Psi_{\gamma,v,\epsilon}(\omega)=\frac{\overline{\widehat{\eta}_v(\omega)}}{\omega^{(1+\gamma)/2}}\mathbb{1}\{\omega\in A_\epsilon\}.$$

Notice that $\Psi_{\gamma,v,\epsilon}$ is bounded and compactly supported and thus lies in $L^p(\mathbb{R})$ for any p, and so in particular

$$\Psi_{\gamma,v,\epsilon}\in L^1(\mathbb{R})\cap L^2(\mathbb{R}),$$

which ensures that

$$\widehat{\Psi}_{\gamma,v,\epsilon} \in L^\infty(\mathbb{R}) \cap L^2(\mathbb{R}).$$

Finally, let us write

$$\Psi_{\gamma,v}(\omega) = \lim_{\epsilon \to 0} \Psi_{\gamma,v,\epsilon}(\omega) = \frac{\overline{\widehat{\eta}_v(\omega)}}{\omega^{(1+\gamma)/2}} \mathbb{1}\{\omega > 0\}$$

for every ω. We now show that $\Psi_{\gamma,v} \in L^2(\mathbb{R})$ provided $M_\gamma(\mu + \nu) < \infty$, which is assumed throughout. Let $(X, Y) \sim \mu \otimes \nu$. We have

$$\begin{aligned}\int_{\mathbb{R}} |\Psi_{\gamma,v}(\omega)|^2 \mathrm{d}\omega &= \int_0^\infty \frac{|\widehat{\eta}_v(\omega)|^2}{w^{1+\gamma}} \mathrm{d}w \\ &= \int_0^\infty \frac{(\mathbb{E}[\cos\langle \omega, X\rangle - \cos\langle \omega, Y\rangle])^2 + (\mathbb{E}[\sin\langle \omega, X\rangle - \sin\langle \omega, Y\rangle])^2}{\omega^{1+\gamma}} \mathrm{d}\omega.\end{aligned}$$

Using the inequality $(a-b)^2 \leq 2(a-1)^2 + (b-1)^2$, $\forall\, a, b \in \mathbb{R}$ for the cos term, the inequality $(a+b) \leq 2a^2 + 2b^2$, $\forall\, a, b \in \mathbb{R}$ for the sin term, and applying Jensen's inequality to take the expectation outside, we can conclude that $\Psi_{\gamma,v} \in L^2(\mathbb{R})$ by Lemma D.1.3. Thus, by the dominated convergence theorem

$$\|\Psi_{\gamma,v,\epsilon} - \Psi_{\gamma,v}\|_2 \to 0$$

as $\epsilon \to 0$. Then, by Parseval's identity

$$\left\|\widehat{\Psi}_{\gamma,v,\epsilon} - \widehat{\Psi}_{\gamma,v}\right\|_2 \to 0 \tag{D.2.2}$$

as $\epsilon \to 0$. It is well know that convergence in $L^2(\mathbb{R})$ implies that there exists a subsequence $\{\epsilon_n\}_{n=1}^\infty$ with $\epsilon_n \to 0$ and $\widehat{\Psi}_{\gamma,v,\epsilon} \to \widehat{\Psi}_{\gamma,v}$ almost everywhere.[2] Therefore, by passing to this subsequence, it follows that

$$\int_{\mathbb{R}} \psi_\gamma(x-b) \mathrm{d}\eta_v(x) = C_{\psi_\gamma} \begin{cases} \mathsf{im}\left(\widehat{\Psi}_{\gamma,v}(b)\right) & \text{if } \gamma = 1 \\ \Re\left(\widehat{\Psi}_{\gamma,v}(b)\right) & \text{if } \gamma \neq 1 \end{cases}$$

for $\sigma \otimes \text{Leb}$-almost every $(b, v) \in \mathbb{R} \times \mathbb{S}^{d-1}$. Note that since $\eta_v(\omega) \in \mathbb{R}$,

$$\begin{aligned}\Re\left(\widehat{\Psi}_{\gamma,v}(b)\right) &= \frac{\widehat{\Psi}_{\gamma,v}(b) + \overline{\widehat{\Psi}_{\gamma,v}(b)}}{2} \\ &= \frac{1}{2}\int_0^\infty \left(\frac{\widehat{\eta}_v(\omega)}{\omega^{(1+\gamma)/2}} e^{ib\omega} + \frac{\widehat{\eta}_v(-\omega)}{\omega^{(1+\gamma)/2}} e^{-ib\omega}\right) \mathrm{d}\omega \\ &= \frac{1}{2}\int_{-\infty}^\infty \frac{\widehat{\eta}_v(\omega)\,\mathrm{sign}(\omega)}{|\omega|^{(1+\gamma)/2}} e^{ib\omega} \mathrm{d}\omega \\ &= \mathcal{F}\left[\frac{\widehat{\eta}_v(\omega)\,\mathrm{sign}(\omega)}{2|\omega|^{(1+\gamma)/2}}\right](-b),\end{aligned} \tag{D.2.3}$$

[2] We could also conclude this by Carleson's theorem.

$$\begin{aligned}
\mathsf{im}\left(\widehat{\Psi}_{\gamma,v}(b)\right) &= \frac{\widehat{\Psi}_{\gamma,v}(b) - \overline{\widehat{\Psi}_{\gamma,v}(b)}}{2i} \\
&= \frac{1}{2i}\int_0^\infty \left(\frac{\overline{\widehat{\eta}_v(\omega)}}{\omega^{(1+\gamma)/2}} e^{-ib\omega} - \frac{\overline{\widehat{\eta}_v(-\omega)}}{\omega^{(1+\gamma)/2}} e^{ib\omega} \right) \mathrm{d}\omega \\
&= \frac{1}{2i}\int_{-\infty}^\infty \frac{\overline{\widehat{\eta}_v(\omega)}}{|\omega|^{(1+\gamma)/2}} \operatorname{sign}(\omega) e^{ib\omega} \mathrm{d}\omega \\
&= \mathcal{F}\left[\frac{\overline{\widehat{\eta}_v(\omega)} \operatorname{sign}(\omega)}{2i|\omega|^{(1+\gamma)/2}} \right](-b). \qquad \text{(D.2.4)}
\end{aligned}$$

Plugging (D.2.3) and (D.2.4) into (D.2.1), by Parseval's identity (implicitly using that $\Psi_{\gamma,v} \in L^2(\mathbb{R})$), we obtain

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \left[\mathbb{E}\psi_\gamma(\langle X, v\rangle - b) - \mathbb{E}\psi_\gamma(\langle Y, v\rangle - b) \right]^2 \mathrm{d}b\mathrm{d}\sigma(v) \\
&\qquad = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \psi_\gamma(x-b)\mathrm{d}\eta_v(x) \right)^2 \mathrm{d}b\mathrm{d}\sigma(v), \\
&\qquad = 2\pi C_{\psi_\gamma}^2 \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \frac{|\widehat{\eta}_v(\omega)|^2}{4|\omega|^{1+\gamma}} \mathrm{d}w\mathrm{d}\sigma(v) \\
&\qquad = \pi C_{\psi_\gamma}^2 \int_{\mathbb{S}^{d-1}} \int_0^\infty \frac{|\widehat{\eta}_v(\omega)|^2}{|\omega|^{1+\gamma}} \mathrm{d}w\mathrm{d}\sigma(v) \\
&\qquad = \pi C_{\psi_\gamma}^2 \int_{\mathbb{R}^d} \frac{|\mathcal{F}[\mu-\nu](\omega)|^2}{\|\omega\|^{d+\gamma}} \mathrm{d}\omega,
\end{aligned}$$

where the last step uses a polar change of variable. The result follows after comparing with Proposition 5.2.3.

D.3 Proof of Proposition 5.2.6

Let $\mathcal{S}(\mathbb{R}^d)$ be the Schwartz space and $\mathcal{S}'(\mathbb{R}^d)$ be the space of all tempered distributions on $\mathbb{R}^d$. Let $\tau = \mu - \nu$ and $s = (d+\gamma)/2$. First, note that

$$\int \frac{K_s(x)\mathrm{d}x}{(1+\|x\|^2)^d} < \infty,$$

so by [138, Theorem 0.10] we have $K_s \in \mathcal{S}'(\mathbb{R}^d)$. By [138, Theorem 0.12], since $K_s \in \mathcal{S}'(\mathbb{R}^d)$ and τ has compact support,

$$\widehat{I_s f} = \widehat{K_s * \tau} = \widehat{K_s}\widehat{\tau}.$$

By Plancherel's identity,

$$(2\pi)^{\frac{d}{2}} \|I_s \tau\|_2 = \|\widehat{I_s \tau}\|_2 = \|\widehat{K_s}\widehat{\tau}\|_2 = \frac{1}{\sqrt{F_\gamma(d)}} \mathcal{E}_\gamma(\mu,\nu),$$

where the last equality follows from Proposition 5.2.3.

D.4 Proof of Theorem 5.3.3 and Proposition 5.3.4

In this section we prove both Proposition 5.3.3 and Proposition 5.3.4. To do so, we give two constructions. The first one, presented in Appendix D.4.1, only applies in one dimension and gives optimal results. The second construction is given in Appendix D.4.2 applies in all dimensions, but loses a polylogarithmic factor.

Notation: Abusing notation, in what follows we write $\mathcal{E}_\gamma(f,g)$ and $\overline{d_H}(f,g)$ even when f and g are not necessarily probability measures or probability densities. We will also write $\|f\|_{t,2} = \||\cdot|^t \widehat{f}\|_2$ for potentially negative exponents $t \in \mathbb{R}$. Note that $\mathcal{E}_\gamma(f,0) = \sqrt{F_\gamma(d)}\|\widehat{f}\|_{-\frac{d+\gamma}{2},2}$.

D.4.1 The Case $d = 1$

The Lemma below constructs the *difference* of two densities that has favorable properties.

Lemma D.4.1. *Let $f(x) = 1\{|x| \le \pi\}\sin(rx)$ with $r \in \mathbb{Z}$ and write $f_\beta = f * \cdots * f$ for f convolved with itself $\beta - 1$ times, i.e. $f_1 = f, f_2 = f * f$ and so on. Fix an integer $\beta \ge 1$ and let $|t| < \beta$. We have*

$$\|f_\beta\|_{t,2} \asymp r^t, \|f_\beta\|_1 \asymp 1, \text{ and } \overline{d_H}(f_\beta, 0) \asymp \frac{1}{r}, \tag{D.4.1}$$

as $r \to \infty$ where the constants may depend on β, t.

Proof. The intuition for the estimates (D.4.1) is simple: most of the energy of f (and hence f_β) is at frequencies around $|\omega| \approx r$ and thus differentiating t times boosts the L_2-energy by r^t. A simple computation shows $\widehat{f}(\omega) = c\frac{(-1)^r}{i}\frac{r}{\omega^2 - r^2}\sin(\omega\pi)$. Note that because $r \in \mathbb{Z}$ we have $\|\widehat{f}\| \asymp \|\widehat{f_\beta}\| \asymp 1$.

Estimating $\|f_\beta\|_{t,2}$. By definition we have

$$\|f_\beta\|_{t,2}^2 \asymp \int_0^\infty |\widehat{f}(\omega)|^{2\beta}\omega^{2t}\,\mathrm{d}\omega \asymp \int_0^\infty \frac{r^{2\beta}}{(\omega^2 - r^2)^{2\beta}}\omega^{2t}\sin^{2\beta}(\omega\pi)\mathrm{d}\omega.$$

We decompose the integral into three regimes:

1. $\underline{\omega < r/2}$: here $(\omega^2 - r^2) \asymp r^2$ and thus

$$\int_0^{r/2}(\dots) \asymp r^{-2\beta}\int_0^{r/2}\omega^{2t}\sin^{2\beta}(\omega\pi) \lesssim r^{2t}$$

   by Lemma D.1.8.

2. $\underline{\omega > 3r/2}$: here $(\omega^2 - r^2) \asymp \omega^2$ and thus

$$\int_{3r/2}^\infty(\dots) \asymp r^{2\beta}\int_{3r/2}^\infty \frac{\sin^{2\beta}(\omega\pi)\omega^{2t}}{\omega^{4\beta}} \asymp r^{2\beta}r^{2t-4\beta+1} = r^{1+2t-2\beta} \ll r^{2t}.$$

3. $\omega \in [r/2, 3r/2]$: here $(\omega^2 - r^2) \asymp yr$, where $y = \omega - r$. Note also $\sin(\omega\pi) = \sin(r\pi + y\pi) = (-1)^r \sin(y\pi)$, and $\omega \asymp r$. Thus

$$\int_{r/2}^{3r/2} (\dots) \mathrm{d}\omega = \int_{-r/2}^{r/2} (\dots) \mathrm{d}y \asymp r^{2\beta} \int_{-r/2}^{r/2} \sin^{2\beta}(y\pi) r^{2t} (yr)^{2\beta} \mathrm{d}y$$
$$\asymp r^{2t} \int_{\mathbb{R}} \left(\frac{\sin(y\pi)}{y} \right)^{2\beta} \mathrm{d}y \asymp r^{2t}.$$

where the last inequality follows by that the integrand is bounded at 0 and has $y^{-2\beta} \lesssim y^{-2}$ tail.

Estimating $\|f_\beta\|_1$. Follows from $\|f_\beta\|_1 \lesssim \|f_\beta\|_2 \asymp 1$ by the Cauchy-Schwartz inequality and $\|f_\beta\|_1 \geq \|\widehat{f_\beta}\|_\infty \asymp 1$ by the Hausdorff–Young inequality.

Estimating $\overline{d_H}$. We get $\overline{d_H}(f_\beta, 0) \gtrsim \mathcal{E}_1(f_\beta, 0) \asymp \|f_\beta\|_{-1,2} \asymp \frac{1}{r}$ from the first estimate. For the upper bound, note that $\widehat{\mathrm{sign}(x)} = \frac{2}{i\omega}$ and $\overline{d_H}(f_\beta, 0) = \sup_b \frac{1}{2} \int f_\beta(x) \,\mathrm{sign}(x - b) \mathrm{d}x$, so by Plancherel's identity,

$$\overline{d_H}(f_\beta, 0) \lesssim \sup_b \int \left| \widehat{f_\beta}(\omega) \frac{e^{ib\omega}}{\omega} \right| \mathrm{d}\omega \lesssim \int_0^\infty \frac{r^\beta}{(\omega^2 - r^2)^\beta} \omega^{-1} \sin^\beta(\omega\pi).$$

The fact that the above is $O(1/r)$ follows analogously to the proof of our bound on $\|f_\beta\|_{t,2}$ so we omit it. This concludes our proof.

□

Proof of Proposition 5.3.3 and Proposition 5.3.4 for $d = 1$. We now turn to showing tightness in one dimension, utilizing the density difference constructed in Lemma D.4.1. Given a value of the smoothness $\beta > 0$, set $\overline{\beta} = \lceil \beta \rceil + 1$ and let $f_{\overline{\beta}}$ be as in Lemma D.4.1 with $r = \epsilon^{-1/\beta}$ for some $\epsilon \in (0, 1)$. Let p_0 be a smooth, compactly supported density with $\inf_{x \in [-\pi, \pi]} p_0(x) > 0$. Define

$$p_\epsilon(x) = p_0(x) + \epsilon f_{\overline{\beta}}(\overline{\beta} x)/2 \qquad \text{and} \qquad q_\epsilon(x) = p_0(x) - \epsilon f_{\overline{\beta}}(\overline{\beta} x)/2.$$

Clearly both p_ϵ, q_ϵ are compactly supported probability densities for sufficiently small ϵ, since $\|f_{\overline{\beta}}\|_\infty < \infty$ and is supported on $[-\overline{\beta}\pi, \overline{\beta}\pi]$. By Lemma D.4.1, for each $\gamma \in (0, 2)$ the two densities satisfy

$$\|p_\epsilon - q_\epsilon\|_1 \asymp \epsilon, \ \|p_\epsilon\|_{\beta,2} \asymp \|q_\epsilon\|_{\beta,2} \asymp 1, \ \mathcal{E}_\gamma(p_\epsilon, q_\epsilon) \asymp \|p_\epsilon - q_\epsilon\|_{-(1+\gamma)/2,2} \asymp \epsilon^{\frac{2\beta+\gamma+1}{2\beta}}, \overline{d_H}(p_\epsilon, q_\epsilon) \asymp \epsilon^{\frac{\beta+1}{\beta}}.$$

This proves both Proposition 5.3.3 and Proposition 5.3.4 for $d = 1$. □

D.4.2 The Case $d > 1$

We move on to the case of general dimension. In Appendix D.4.2 we outline our approach. Then, in Appendix D.4.2 we give full details of our construction, following the argument outlined in the prior section.

Overview

For the discussions below, we will assume that the ambient dimension $d \geq 2$. Our construction here is less straightforward than for $d = 1$ in Appendix D.4.1 but shares the same basic idea. Recall that the basic premise is that we want to saturate the Hölder's inequality in Equation (5.3.3), which requires the density difference $f = \mu - \nu$ to have Fourier transform be (almost) supported on a sphere. For $d = 1$ we took f to be a pure sinusoid. However, of course such f is not compactly supported and that is why we multiplied the sinusoid by a rectangle (and then convolved many times to gain smoothness), which served as a mollifier.

For $d > 1$ let us attempt to follow the same strategy and take

$$f_r(x) = g_r(x)h(x)\,,$$

where $r > 0$ is a parameter, h is some compactly supported smooth mollifier and $g_r(x)$ is defined implicitly via

$$\widehat{g}_r(\omega) = r^{(1-d)/2}\delta(\|\omega\| - r),$$

where here and below we denote, a bit informally, by $\delta(\|\cdot\| - r)$ a distribution that integrates any smooth compactly supported function ϕ as follows:

$$\int_{\mathbb{R}^d} \phi(\omega)\delta(\|\omega\| - r)\mathrm{d}\omega =: r^{d-1}\int_{\mathbb{R}^d} \phi(r\omega)\mathrm{d}\sigma(\omega) = \frac{2\pi^{d/2}r^{d-1}}{\Gamma(\frac{d}{2})}\mathbb{E}_{\phi(rX)}[\,]\,,$$

where σ is the unnormalized surface measure of $\mathbb{S}^{d-1}$ and X is a random vector uniformly distributed on $\mathbb{S}^{d-1}$. Explicit computation shows

$$\begin{aligned} g_r(x) = \mathcal{F}^{-1}[\widehat{g}_r](x) &= \frac{\sqrt{r}}{(2\pi)^d r^{d/2}}\int_{\mathbb{R}^d} e^{i\langle\omega,x\rangle}\delta(\|\omega\| - r)\mathrm{d}\omega \\ &= \frac{\sqrt{r}}{(2\pi)^{d/2}}\,\|x\|^{1-d/2}\,J_{d/2-1}(\|rx\|), \end{aligned}$$

where J_ν denotes Bessel functions of the first kind of order ν. Notice that g is spherically symmetric and real-valued (some further properties of it are collected below in Lemma D.1.10).

Note that $|g_r(x)| = O(1)$ as $r \to \infty$ for any fixed $x \neq 0$ (Lemma D.1.10), while at the origin we have $|g_r(0)| = \Omega(r^{(d-1)/2})$, which follows from the series expansion of the Bessel function given in for example [203, Section 3.1-3.11]. This causes an issue for $d > 1$, as g_r is too large at the origin as $r \to \infty$ compared to its tails, which makes it difficult to use it as the difference between two probability densities. Hence, we choose our mollifier h to be supported on an annulus instead of on a ball. In addition, it will also be convenient for it to have a super-polynomially decaying Fourier transform, i.e.

$$|\widehat{h}(w)| \leq H(\|w\|) \triangleq C\exp\left(-\frac{c\|w\|}{\log(\|w\|+2)^2}\right) \qquad \forall w \in \mathbb{R}^d\,.$$

The existence of the desired function h is proven Lemma D.4.2.

Note that all of the Fourier energy of g_r lies at frequencies $\|\omega\| = r$ by construction. However, after multiplying by h the energy spills over to adjacent frequencies as well and we

need to estimate the amount of the spill. Due to the fast decay of $\widehat{h}$ we will show, roughly, the following estimates on the behavior of $\widehat{f_r}$:

$$|\widehat{f_r}(\omega)| \lesssim \tilde{O}(r^{(1-d)/2})1\{\|\omega - r\| \le \log^2(r)\} + r^{(d-1)/2} H(\max(\|\omega\| - r, \log^2 r)), \quad \text{and}$$
$$|\widehat{f_r}(\omega)| \lesssim r^{(d-1)/2}\|\omega\|$$

as $r \to \infty$. Note that the first bound above is super-polynomially decaying in both $\|\omega\|$ and r, which allows us to show that

$$\|f_r\|_{t,2} \le \tilde{O}(r^t)$$

for $t > -\frac{d+2}{2}$, recalling the notation $\|f\|_{t,2} = \||\cdot|^t \widehat{f}\|_2$. A direct calculation will also show

$$\|f_r\|_1 \asymp \|f_r\|_\infty \asymp \|f_r\|_2 \asymp 1\,.$$

For a desired total-variation separation ϵ, we will set $\mu - \nu = \epsilon f_r$ and choose $r = \epsilon^{-1/\beta}$ to ensure that $\epsilon\|f_r\|_{\beta,2} = \tilde{O}(1)$. For the energy distance between μ and ν these choices yield

$$\mathcal{E}_\gamma(\mu,\nu) \asymp \|\epsilon f_{\epsilon^{-1/\beta}}\|_{-\frac{d+\gamma}{2},2} = \tilde{O}(\epsilon^{1+\frac{d+\gamma}{2\beta}}) = \tilde{O}(\mathsf{TV}^{\frac{d+2\beta+\gamma}{2\beta}})\,,$$

as required.

We now proceed to rigorous details.

The construction

First, we must construct the mollifier h with the properties outlined in Appendix D.4.2. Recall that a function f is radial (also known as spherically symmetric) if its value at $x \in \mathbb{R}^d$ depends only on $\|x\|$. In other words, $f(x) = f(y)$ holds for all $x, y \in \mathbb{R}^d$ with $\|x\| = \|y\|$.

Lemma D.4.2. *There exists a compactly supported radial Schwartz function h, and a positive sequence $\{r_n\}_{n=1}^\infty$ satisfying $r_n = \Theta(n)$, such that*

$$\operatorname{supp}(h) \subset \mathbb{B}(0,1), \tag{D.4.2}$$
$$\operatorname{supp}(h) \subset \mathbb{R}^d \setminus \mathbb{B}(0,r_0), \tag{D.4.3}$$
$$|\widehat{h}(w)| \le C\exp\left(-\frac{c\|w\|}{\log(\|w\|+2)^2}\right) \quad \textit{for all } w \in \mathbb{R}^d, \textit{ and} \tag{D.4.4}$$
$$\widehat{h}(r_n u) = 0 \quad \textit{for all } u \in \mathbb{S}^{d-1}, \tag{D.4.5}$$

for universal constants $C, c, r_0 > 0$.

Proof. First, let h_0 be as constructed in Lemma D.1.11, which already satisfies Equation (D.4.2) and Equation (D.4.4). To address the other two requirements, we modify h_0 by convolving it with two additional terms:

$$h(x) := (A_0(\cdot) * h_0(8\cdot) * \rho_0(\cdot))(x),$$

where A_0 and ρ_0 aim to address Equation (D.4.3) and Equation (D.4.5), respectively, and are defined as

$$A_0(x) = \exp\left(-\frac{1}{1/64 - (\|x\| - 1/2)^2}\right)\mathbb{1}\{\|x\| \in (3/8, 5/8)\}, \quad \rho_0(x) = \mathbb{1}\{\|x\| < 1/8\}.$$

Before proceeding, note that clearly h is a radial Schwartz function. Let us now verify that h indeed satisfies the four requirements. Note that A_0 is an "annulus" supported on $\mathbb{B}(0,5/8)\backslash\mathbb{B}(0,3/8)$, and both $h_0(8\cdot)$ and ρ_0 are supported on $\mathbb{B}(0,1/8)$. Therefore, $\mathrm{supp}(h) \subset \mathbb{B}(0,7/8)\backslash\mathbb{B}(0,1/8)$, which implies Equations (D.4.2) and (D.4.3). We now turn to the other two conditions in Fourier space. Note that

$$\widehat{h}(w) = (1/8)^d \cdot \widehat{A}_0(w) \cdot \widehat{h}_0(w/8) \cdot \widehat{\rho}_0(w).$$

From Item 3 of Lemma D.1.10 we know that

$$\mathcal{F}[\mathbb{1}\{\|\cdot\| < 1\}](w) = \left(\frac{2\pi}{\|w\|}\right)^{\frac{d}{2}} J_{\frac{d}{2}}(\|w\|).$$

Hence, by Item 2 of Lemma D.1.10, the function $\widehat{\rho}_0(w) = (1/8)^d \mathcal{F}[\mathbb{1}\{\|\cdot\| < 1\}](w/8)$ has infinitely many zeros near the values of $\|w\| = 8(2n\pi + \frac{(d+1)\pi}{4})$ for sufficiently large $n \in \mathbb{Z}^+$, which implies Equation (D.4.5).

Finally, for Equation (D.4.4), note that since both A_0 and ρ_0 are Schwartz functions, so are their Fourier transforms $\widehat{A}_0$ and $\widehat{\rho}_0$ so that

$$|\widehat{h}(w)| \leq (1/8)^d \|\widehat{A}_0\|_\infty \|\widehat{\rho}_0\|_\infty |\widehat{h}_0(w/8)| \lesssim |\widehat{h}_0(w/8)|,$$

concluding the proof. □

Let h be as constructed in Lemma D.4.2, and define

$$f_r = g_r h \tag{D.4.6}$$

for $r > 0$ and $\widehat{g}(\omega) =: r^{(1-d)/2}\delta(\|\omega\| - r)$. Recall from the overview of our construction that we gave in Appendix D.4.2 that f_r is our proposed density difference which we claim (approximately) saturates Hölder's inequality in (5.3.3). The next Lemma records the properties of f_r which will enable us to complete our proof.

Lemma D.4.3. *Let f_r be as in (D.4.6) and let $\{r_n\}_{n=1}^\infty$ be the sequence constructed in Lemma D.4.2. The following hold.*

(i) For all $n \in \mathbb{N}$ we have

$$\int_{\mathbb{R}^d} f_{r_n}(x)dx = 0 \qquad \text{and} \qquad \mathrm{supp}(f_{r_n}) \subset \mathbb{B}(0,1).$$

(ii) We have

$$\|f_{r_n}\|_\infty \asymp \|f_{r_n}\|_2 \asymp \|f_{r_n}\|_1 \asymp 1,$$

hiding constants independent of n.

(iii) For any $t > -\frac{d+2}{2}$ we have

$$\|f_{r_n}\|_{t,2} = O(r_n^t \log^d(r_n))$$

as $n \to \infty$, hiding constants independent of n.

(iv) Recall the definition of ψ_γ from (5.2.6). For any $\gamma \in (0,2)$ we have

$$\sup_{v\in\mathbb{S}^{d-1},b\in\mathbb{R}} \left| \int_{\mathbb{R}^d} \psi_\gamma(\langle x, v\rangle - b) f_{r_n}(x)\mathrm{d}x \right| = O(r_n^{-(d+\gamma)/2} \log(r_n)^d)$$

hiding constants independent of n.

Proof. Let us drop the dependence of r_n to simplify notation.
Showing (i). Note that $\int_{\mathbb{R}^d} f_r(x)\mathrm{d}x = \widehat{f}_r(0)$. Then, $\widehat{f}_r(0) = 0$ follows from the construction of h and g_r. Indeed, $\widehat{g}_r$ is supported on $r\mathbb{S}^{d-1}$ while $\widehat{h}|_{r\mathbb{S}^{d-1}} \equiv 0$. The fact that $\mathrm{supp}(f_r) \subset \mathbb{B}(0,1)$ follows from $\mathrm{supp}(h) \subset \mathbb{B}(0,1)$.

Showing (ii). Since f_r has compact support, we immediately have

$$\|f_r\|_1 \lesssim \|f_r\|_2 \lesssim \|f_r\|_\infty.$$

As h is continuous and supported on the annulus $\{x : r_0 \le \|x\| \le 1\}$ by construction, it suffices to bound g_r on said annulus. Now, for any x with $r_0 \le \|x\| \le 1$, we have by Lemma D.1.10 that

$$g_r(x) \lesssim \sqrt{r}\|x\|^{1-d/2} \frac{1}{\sqrt{r\|x\|}} \lesssim 1,$$

which shows that $\|f_r\|_\infty \lesssim 1$.

We now turn to lower bounding $\|f_r\|_1$. Recall that h is uniformly continuous and nontrivial, hence $\int |h(u^* v)|\mathrm{d}\sigma(v) \neq 0$ for some radius u^*, and thus for all $u \in (u_0, u_1) \subseteq (0,1)$ for some constants u_0, u_1. Using that g_r is spherically symmetric, we compute

$$\begin{aligned}
\|f\|_1 &= \int_{\mathbb{R}^d} |g_r(x)||h(x)|\mathrm{d}x \\
&= \int_0^\infty u^{d-1} g_r(u,0,\dots,0) \int h(uv)\mathrm{d}\sigma(v)\mathrm{d}u \\
&\gtrsim \sqrt{r} \int_{u_0}^{u_1} \left|J_{d/2-1}(ru)\right| \mathrm{d}u \gtrsim 1,
\end{aligned}$$

where the last line follows by (D.1.1) once again.

Showing (iii). Let $0 < s < r$, whose precise value will be set later. For convenience, set $B_s = \{x \in \mathbb{R}^d : \|x\| \le s\}$ and $B_s^c = \mathbb{R}^d \setminus B_s$. Recall that by definition

$$\begin{aligned}
\widehat{f}_r(\omega) &= r^{(1-d)/2} \int_{\mathbb{R}^d} \widehat{h}(\omega + x)\delta(\|x\| - r)\mathrm{d}x \\
&= \underbrace{r^{(1-d)/2} \int_{\mathbb{R}^d} (\widehat{h}\mathbb{1}_{B_s})(\omega + x)\delta(\|x\| - r)\mathrm{d}x}_{I} + \underbrace{r^{(1-d)/2} \int_{\mathbb{R}^d} (\widehat{h}\mathbb{1}_{B_s^c})(\omega + x)\delta(\|x\| - r)\mathrm{d}x}_{II}.
\end{aligned} \tag{D.4.7}$$

Let C, c be as in Lemma D.4.2, and $H(x) = C\exp(-c\|x\|/\log^2(\|x\|+2))$. Note that $\|\widehat{h}\|_\infty \le C$. Therefore, the first term in the decomposition (D.4.7) can be bounded by

$$\begin{aligned}
|I| &\le Cr^{(1-d)/2}\int_{\mathbb{R}^d} \mathbb{1}\{\|\omega+x\|\le s\}\delta(\|x\|-r)\mathrm{d}x \\
&= Cr^{(1-d)/2}\mathbb{1}\{\|\omega\|\in[r-s,r+s]\}\int_{\mathbb{R}^d} \mathbb{1}\{\|\omega+x\|\le s\}\delta(\|x\|-r)\mathrm{d}x \\
&\lesssim r^{(1-d)/2}s^{d-1}\mathbb{1}\{\|\omega\|\in[r-s,r+s]\}.
\end{aligned}$$

The second line uses that if $\|\omega\| \notin [r-s, r+s]$ then the integral becomes zero. The third line uses the fact that the surface area of the intersection of B_s with any sphere of any radius (and the one centered at ω with radius r in particular) is at most $O(s^{d-1})$.

Moving on to the second term, we have

$$|II| = r^{(d-1)/2}\int |(\widehat{h}\mathbb{1}_{B_s^c})(\omega+ru)|\mathrm{d}\sigma(u) \lesssim r^{(d-1)/2}H(\max\{\|\omega\|-r, s\})$$

using that $H:[0,\infty)\to(0,C]$ is decreasing and that $|\widehat{h}(y)\mathbb{1}_{B_s^c}(y)\| \le H(\max\{y,s\})$ for all $y\in\mathbb{R}^d$. Summarizing, we have the pointwise estimate

$$|\widehat{f_r}(\omega)| \lesssim r^{(1-d)/2}s^{d-1}\mathbb{1}\{\|\omega\|\in[r-s,r+s]\} + r^{(d-1)/2}H(\max\{\|\omega\|-r,s\}) \tag{D.4.8}$$

for all $\omega\in\mathbb{R}^d$ and $0<s<r$.

We now show that f_r is Lipschitz continuous. Recall from the construction of h (Lemma D.4.2) that $h|_{r_n\mathbb{S}^{d-1}} \equiv 0$. Then, we observe that for any $\omega\in\mathbb{R}^d$

$$\begin{aligned}
|\widehat{f_r}(\omega)| &= r^{(d-1)/2}\left|\int \widehat{h}(\omega+ru)\mathrm{d}\sigma(u)\right| \\
&= r^{(d-1)/2}\left|\int \{\widehat{h}(\omega+ru)-\widehat{h}(ru)\}\mathrm{d}\sigma(u)\right| \\
&= r^{(d-1)/2}\|\widehat{h}\|_{\mathrm{Lip}}\frac{2\pi^{d/2}\|\omega\|}{\Gamma(\frac{d}{2})} \\
&\lesssim r^{(d-1)/2}\|\omega\|,
\end{aligned} \tag{D.4.9}$$

where we use that $\widehat{h}$ is Schwartz by construction, and thus has finite Lipschitz constant $\|\widehat{h}\|_{\mathrm{Lip}}$.

With (D.4.8) and (D.4.9) in hand we can proceed to bounding the norm of f_r. Let $s = D\log(r)^2$ for a large constant D independent of r, and assume that r is large enough so

that $s < r/2$. Also set $\theta > 0$, whose precise value is specified later. We have

$$\begin{aligned}
\|f_r\|_{t,2}^2 &= \int_{\mathbb{R}^d} \|\omega\|^{2t} |\widehat{f_r}(\omega)|^2 \mathrm{d}\omega \\
&\overset{(D.4.9)}{\lesssim} r^{d-1} \int_{\|\omega\| \le r^{-\theta}} \|\omega\|^{2t+2} \mathrm{d}\omega + \int_{\|\omega\| > r^{-\theta}} \|\omega\|^{2t} |\widehat{f_r}(\omega)|^2 \mathrm{d}\omega \\
&\overset{(D.4.8)}{\lesssim} r^{d-1-\theta(2t+d)} \\
&\qquad + r^{1-d} \log(r)^{2(d-1)} \int_{\|\omega\| > r^{-\theta}} \|\omega\|^{2t} \mathbb{1}\{\|\omega\| \in [r-s, r+s]\} \mathrm{d}\omega \\
&\qquad + r^{d-1} \int_{\|\omega\| > r^{-\theta}} \|\omega\|^{2t} H^2(\max\{\|\omega\| - r, s\}) \mathrm{d}\omega \\
&\lesssim r^{d-1-\theta(2t+d)} + r^{2t} \log(r)^{2d-1} + r^{d-1} \int_{r^{-\theta}}^{\infty} u^{2t+d-1} H^2(\max\{u-r, s\}) \mathrm{d}u.
\end{aligned}$$

Note that in the derivation above we changed to polar coordinates freely, and that in the second inequality we used the assumption $t > -d/2 - 1$. Setting θ to any positive value greater than $(d-1-2t)/(2t+d)$ ensures that the first term in the final line is $O(r^{2t})$. As for the integral term, we can bound it by

$$\lesssim r^{d-1} H^2(s) \int_{r^{-\theta}}^{2r} u^{2t+d-1} \mathrm{d}u + r^{d-1} \int_{2r}^{\infty} H^2(u/2) \mathrm{d}u \overset{\text{Lemma D.1.9}}{\lesssim} \mathrm{poly}(r) \times H^2(s) + r^{2t}.$$

By taking D large enough (independently of r) in the definition of $s = D \log^2(r)$ we can make also the first term $\mathrm{poly}(r) \times H^2(s)$ less than $O(r^{2t})$, which concludes the proof of *(iii)*.

Showing (iv). The bounds that we develop below are analogous to those given in the proof of *(iii)*. Fix $b \in \mathbb{R}$ and $v \in \mathbb{S}^{d-1}$ and define

$$\dagger := \int_{\mathbb{R}^d} \psi_\gamma(\langle v, x \rangle - b) f_r(x) \mathrm{d}x.$$

Suppose first that $\gamma \neq 1$. Then, using Lemmas D.1.4 and D.1.5, we know by dominated convergence that

$$\begin{aligned}
\dagger &= \int_{\mathbb{R}^d} \lim_{\epsilon \to 0} \int_{\epsilon}^{1/\epsilon} C_{\psi_\gamma} \frac{\cos(t(\langle v, x \rangle - b)) - \mathbb{1}\{\gamma > 1\}}{t^{(1+\gamma)/2}} f_r(x) \mathrm{d}t \mathrm{d}x \\
&= C_{\psi_\gamma} \lim_{\epsilon \to 0} \int_{\epsilon}^{1/\epsilon} \Re \left\{ \int_{\mathbb{R}^d} \frac{e^{it(\langle v, x \rangle - b)} f_r(x)}{t^{(1+\gamma)/2}} \mathrm{d}x \right\} \mathrm{d}t \\
&= C_{\psi_\gamma} \lim_{\epsilon \to 0} \int_{\epsilon}^{1/\epsilon} \frac{\cos(tb) \widehat{f_r}(tv)}{t^{(1+\gamma)/2}} \mathrm{d}t.
\end{aligned}$$

Similarly, for $\gamma = 1$ we can compute

$$\dagger = C_{\psi_\gamma} \lim_{\epsilon \to 0} \int_{\epsilon}^{1/\epsilon} \frac{\sin(-tb) \widehat{f_r}(tv)}{t} \mathrm{d}t.$$

In either case, we have $|\dagger| \lesssim \int_0^\infty |\widehat{f}_r(tv)|/t^{(1+\gamma)/2}\mathrm{d}t$.

Let $s = D\log^2(r)$ for large D independent of r as in the proof of (iii), and let $\theta > 0$ whose precise value is specified later. Assuming that r is large enough so that $s < r/2$, for any $\gamma \in (0,2)$ we have

$$\begin{aligned}
|\dagger| &\leq \int_0^{r^{-\theta}} \frac{|\widehat{f}_r(tv)|}{t^{(1+\gamma)/2}}\mathrm{d}t + \int_{r^{-\theta}}^\infty \frac{|\widehat{f}_r(tv)|}{t^{(1+\gamma)/2}}\mathrm{d}t \\
&\overset{(D.4.9)}{\lesssim} r^{(d-1)/2}\int_0^{r^{-\theta}} t^{(1-\gamma)/2}\mathrm{d}t + \int_{r^{-\theta}}^\infty \frac{|\widehat{f}_r(tv)|}{t^{(1+\gamma)/2}}\mathrm{d}t \\
&\overset{(D.4.8)}{\lesssim} r^{\frac{d-1}{2}-\theta\frac{3-\gamma}{2}} \\
&\quad + \int_{r^{-\theta}}^\infty \frac{1}{t^{(1+\gamma)/2}}\left(r^{(1-d)/2}s^{d-1}\mathbb{1}\{t\in[r-s,r+s]\} + r^{(d-1)/2}H(\max\{t-r,s\})\right)\mathrm{d}t \\
&\lesssim r^{\frac{d-1}{2}-\theta\frac{3-\gamma}{2}} + r^{-(d+\gamma)/2}\log^d(r) + H(s)r^{(d-1)/2}\int_{r^{-\theta}}^{2r}\frac{\mathrm{d}t}{t^{(1+\gamma)/2}} + \int_{2r}^\infty \frac{H(t/2)}{t^{(1+\gamma)/2}}\mathrm{d}t \\
&\overset{\text{Lemma D.1.9}}{\lesssim} r^{\frac{d-1}{2}-\theta\frac{3-\gamma}{2}} + r^{-(d+\gamma)/2}\log^d(r) + H(s)\times\mathrm{poly}(r) + r^{-100d}, \qquad \text{(D.4.10)}
\end{aligned}$$

Set θ to any value greater than $(2d+\gamma-1)/(3-\gamma)$, which ensures that the first term in (D.4.10) is $O(r^{-(d+\gamma)/2})$. By taking D large enough in the definition of $s = D\log^2(r)$, we can make $H(s)$ smaller than any polynomial in r, which ensures that the third term in (D.4.10) is also $O(r^{-(d+\gamma)/2})$. We thus obtain the final bound $|\dagger| \lesssim r^{-(d+\gamma)/2}\log^d(r)$, concluding our proof. □

Proof of Theorem 5.3.3 and Proposition 5.3.4 for $d > 1$. Using the functions $\{f_{r_n}\}_{n=1}^\infty$ we constructed in Lemma D.4.3, we are ready to prove Propositions 5.3.3 and 5.3.4 for $d > 1$.

Let p_0 be a compactly supported probability density with $\inf_{\|x\|\leq 1} p_0(x) > 0$. Fix the smoothness $\beta > 0$. Given any desired total variation separation $\epsilon \in (0,1)$, we can find $n_0 \in \mathbb{N}$ such that $\epsilon^{-1/\beta} \asymp r_{n_0}$, where we hide an ϵ-independent multiplicative constant. Define

$$p_\epsilon = p_0 + \epsilon f_{r_{n_0}}/2 \qquad \text{and} \qquad q_\epsilon = p_0 - \epsilon f_{r_{n_0}}/2.$$

Clearly p_ϵ and q_ϵ are compactly supported probability densities for all small enough ϵ. Moreover, by Lemma D.4.3 they satisfy

$$\begin{aligned}
\|p_\epsilon - q_\epsilon\|_1 &\asymp \epsilon \qquad \text{and} \qquad \|p_\epsilon\|_{\beta,2} \asymp \|q_\epsilon\|_{\beta,2} \asymp 1 \qquad \text{and} \\
\mathcal{E}_\gamma(p_\epsilon, q_\epsilon) &\lesssim \epsilon^{\frac{2\beta+d+\gamma}{2\beta}}\log(1/\epsilon)^d \qquad \text{and} \qquad \overline{d_H}(p_\epsilon, q_\epsilon) \lesssim \epsilon^{\frac{2\beta+d+1}{2\beta}}\log(1/\epsilon)^d
\end{aligned}$$

for all fixed $\gamma \in (0,2)$. This concludes our proof. □

D.5 Proof of Proposition 5.5.1

Proof. Note that $\overline{d_H} = T_{d,0}$. Let p_ϵ, q_ϵ be the compactly supported densities constructed in the proof of Proposition 5.3.3 in the general dimensional case. Then by construction

$$\epsilon \asymp \mathsf{TV}(p_\epsilon, q_\epsilon) \asymp \|p_\epsilon - q_\epsilon\|_2 \quad \text{and} \quad \|p_\epsilon\|_{\beta,2} + \|q_\epsilon\|_{\beta,2} \lesssim 1 \quad \text{and} \quad \overline{d_H}(p_\epsilon, q_\epsilon) \lesssim \epsilon^{\frac{2\beta+d+1}{\beta}}\log(1/\epsilon)^d.$$

Write $p_{\epsilon,n}$ and $q_{\epsilon,n}$ for the empirical measures of p_ϵ and q_ϵ respectively, based on n i.i.d. observations each. By the triangle inequality we have

$$\begin{aligned}\mathbb{E}\overline{d_H}(p_{\epsilon,n}, q_{\epsilon,n}) &\leq \mathbb{E}\overline{d_H}(p_{\epsilon,n}, p_\epsilon) + \overline{d_H}(p_\epsilon, q_\epsilon) + \mathbb{E}\overline{d_H}(q_\epsilon, q_{\epsilon,n}) \\ &\overset{\text{Lemma 5.4.3}}{\lesssim} 1/\sqrt{n} + \epsilon^{\frac{2\beta+d+1}{2\beta}} \log(1/\epsilon)^d.\end{aligned}$$

This completes the proof. □

References

[1] Georges Aad, Tatevik Abajyan, B Abbott, J Abdallah, S Abdel Khalek, Ahmed Ali Abdelalim, R Aben, B Abi, M Abolins, OS AbouZeid, et al. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physics Letters B* 716.1 (2012), pp. 1–29.

[2] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, and Ananda Suresh. "Competitive classification and closeness testing". In: *Conference on Learning Theory*. JMLR Workshop and Conference Proceedings. 2012, pp. 22–1.

[3] Jayadev Acharya, Hirakendu Das, Alon Orlitsky, Shengjun Pan, and Narayana P Santhanam. "Classification using pattern probability estimators". In: *2010 IEEE International Symposium on Information Theory*. IEEE. 2010, pp. 1493–1497.

[4] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. "Optimal probability estimation with applications to prediction and classification". In: *Conference on Learning Theory*. PMLR. 2013, pp. 764–796.

[5] Claire Adam-Bourdarios, Glen Cowan, Cécile Germain, Isabelle Guyon, Balázs Kégl, and David Rousseau. "The Higgs boson machine learning challenge". In: *NIPS 2014 workshop on high-energy physics and machine learning*. PMLR. 2015, pp. 19–55.

[6] Sea Agostinelli, John Allison, K al Amako, John Apostolakis, H Araujo, Pedro Arce, Makoto Asai, D Axen, Swagato Banerjee, GJNI Barrand, et al. "GEANT4—a simulation toolkit". In: *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303.

[7] Justin Alsing, Tom Charnock, Stephen Feeney, and Benjamin Wandelt. "Fast likelihood-free cosmology with neural density estimators and active learning". In: *Monthly Notices of the Royal Astronomical Society* 488.3 (2019), pp. 4440–4458.

[8] Johan Alwall, Pavel Demin, Simon De Visscher, Rikkert Frederix, Michel Herquet, Fabio Maltoni, Tilman Plehn, David L Rainwater, and Tim Stelzer. "MadGraph/MadEvent v4: the new web generation". In: *Journal of High Energy Physics* 2007.09 (2007), p. 028.

[9] Anders Andreassen, Ilya Feige, Christopher Frye, and Matthew D Schwartz. "JUNIPR: a framework for unsupervised machine learning in particle physics". In: *The European Physical Journal C* 79 (2019), pp. 1–24.

[10] Ery Arias-Castro, Bruno Pelletier, and Venkatesh Saligrama. "Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension". In: *Journal of Nonparametric Statistics* 30.2 (2018), pp. 448–471.

[11] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR. 2017, pp. 214–223.

[12] Yu Bai, Tengyu Ma, and Andrej Risteski. "Approximability of discriminators implies diversity in GANs". In: *arXiv preprint arXiv:1806.10586* (2018).

[13] Sivaraman Balakrishnan and Larry Wasserman. "Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates". In: *The Annals of Statistics* 47.4 (2019), pp. 1893–1927.

[14] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. "Searching for exotic particles in high-energy physics with deep learning". In: *Nature communications* 5.1 (2014), pp. 1–9.

[15] Keith Ball. "Eigenvalues of Euclidean distance matrices". In: *Journal of Approximation Theory* 68.1 (1992), pp. 74–82.

[16] Ziv Bar-Yossef. *The complexity of massive data set computations*. University of California, Berkeley, 2002.

[17] Alexander Barg and G David Forney. "Random codes: Minimum distances and error exponents". In: *IEEE Transactions on Information Theory* 48.9 (2002), pp. 2568–2573.

[18] Ludwig Baringhaus and Carsten Franz. "On a new multivariate two-sample test". In: *Journal of multivariate analysis* 88.1 (2004), pp. 190–206.

[19] Tugkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. "Testing random variables for independence and identity". In: *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*. IEEE. 2001, pp. 442–451.

[20] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White. "Testing that distributions are close". In: *Proceedings 41st Annual Symposium on Foundations of Computer Science*. IEEE. 2000, pp. 259–269.

[21] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White. "Testing closeness of discrete distributions". In: *Journal of the ACM (JACM)* 60.1 (2013), pp. 1–25.

[22] Mark A Beaumont. "Approximate bayesian computation". In: *Annual review of statistics and its application* 6 (2019), pp. 379–403.

[23] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. "Mutual information neural estimation". In: *International conference on machine learning*. PMLR. 2018, pp. 531–540.

[24] Giorgio Bertorelle, Andrea Benazzo, and S Mona. "ABC as a flexible framework to estimate demography over space and time: some cons, many pros". In: *Molecular ecology* 19.13 (2010), pp. 2609–2625.

[25] Bhaswar Bhattacharya and Gregory Valiant. "Testing closeness with unequal sized samples". In: *Advances in Neural Information Processing Systems* 28 (2015).

[26] Lucien Birgé. "On estimating a density using Hellinger distance and some other strange facts". In: *Probability theory and related fields* 71.2 (1986), pp. 271–291.

[27] Lucien Birgé. "Robust tests for model selection". In: *From probability to statistics and back: high-dimensional models and processes–A Festschrift in honor of Jon A. Wellner* (2013), pp. 47–64.

[28] Lucien Birgé. *Sur un théoreme de minimax et son application aux tests*. Univ. de Paris-Sud, Dép. de Mathématique, 1979.

[29] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.

[30] Parham Boroumand et al. "Universal Neyman-Pearson Classification with a Known Hypothesis". In: *arXiv preprint arXiv:2206.11700* (2022).

[31] Wacha Bounliphone, Eugene Belilovsky, Matthew B. Blaschko, Ioannis Antonoglou, and Arthur Gretton. "A Test of Relative Similarity For Model Selection in Generative Models". In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2016. URL: http://arxiv.org/abs/1511.04581.

[32] Johann Brehmer, Gilles Louppe, Juan Pavez, and Kyle Cranmer. "Mining gold from implicit models to improve likelihood-free inference". In: *Proceedings of the National Academy of Sciences* 117.10 (2020), pp. 5242–5249.

[33] Jean Bretagnolle and Catherine Huber. "Estimation des densités: risque minimax". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 47 (1979), pp. 119–137.

[34] V Buldygin and K Moskvichova. "The sub-Gaussian norm of a binary random variable". In: *Theory of probability and mathematical statistics* 86 (2013), pp. 33–49.

[35] Clément L Canonne. "A short note on learning discrete distributions". In: *arXiv preprint arXiv:2002.11457* (2020).

[36] Clément L Canonne. "A survey on distribution testing: Your data is big. But is it blue?" In: *Theory of Computing* (2020), pp. 1–100.

[37] Clément L Canonne and Yucheng Sun. "Optimal Closeness Testing of Discrete Distributions Made (Complex) Simple". In: *arXiv preprint arXiv:2204.12640* (2022).

[38] Fernando Castro-Prado, Wenceslao González-Manteiga, Javier Costas, Fernando Facal, and Dominic Edelmann. "Tests for categorical data beyond Pearson: A distance covariance and energy distance approach". In: *arXiv preprint arXiv:2403.12711* (2024).

[39] Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. "Optimal algorithms for testing closeness of discrete distributions". In: *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2014, pp. 1193–1203.

[40] Serguei Chatrchyan, Vardan Khachatryan, Albert M Sirunyan, Armen Tumasyan, Wolfgang Adam, Ernest Aguilo, Thomas Bergauer, M Dragicevic, J Erö, C Fabjan, et al. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Physics Letters B* 716.1 (2012), pp. 30–61.

[41] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. "Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions". In: *arXiv preprint arXiv:2209.11215* (2022).

[42] Xiuyuan Cheng and Alexander Cloninger. "Classification Logit Two-sample Testing by Neural Networks". In: *CoRR* abs/1909.11298 (2019). arXiv: 1909.11298. URL: http://arxiv.org/abs/1909.11298.

[43] Nikolai N Chentsov. "Estimation of unknown probability density based on observations". In: *Dokl. Akad. Nauk SSSR*. Vol. 147. 1962, pp. 45–48.

[44] Julien Chhor and Alexandra Carpentier. "Goodness-of-Fit Testing for Hölder-Continuous Densities: Sharp Local Minimax Rates". In: *arXiv preprint arXiv:2109.04346* (2021).

[45] Julien Chhor and Alexandra Carpentier. "Sharp Local Minimax Rates for Goodness-of-Fit Testing in Large Random Graphs, multivariate Poisson families and multinomials". In: *arXiv preprint arXiv:2012.13766* (2020).

[46] Lynna Chu and Xiongtao Dai. "Manifold energy two-sample test". In: *Electronic Journal of Statistics* 18.1 (2024), pp. 145–166.

[47] Alex Cohen. "Fractal uncertainty in higher dimensions". In: *arXiv preprint arXiv:2305.05022* (2023).

[48] Gennaro Corcella, Ian G Knowles, Giuseppe Marchesini, Stefano Moretti, Kosuke Odagiri, Peter Richardson, Michael H Seymour, and Bryan R Webber. "HERWIG 6: an event generator for hadron emission reactions with interfering gluons (including supersymmetric processes)". In: *Journal of High Energy Physics* 2001.01 (2001), p. 010.

[49] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. "Asymptotic formulae for likelihood-based tests of new physics". In: *The European Physical Journal C* 71.2 (2011), pp. 1–19.

[50] Koby Crammer and Yoram Singer. "On the learnability and design of output codes for multiclass problems". In: *Machine learning* 47 (2002), pp. 201–233.

[51] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. "The frontier of simulation-based inference". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30055–30062.

[52] Katalin Csilléry, Michael GB Blum, Oscar E Gaggiotti, and Olivier François. "Approximate Bayesian computation (ABC) in practice". In: *Trends in ecology & evolution* 25.7 (2010), pp. 410–418.

[53] Niccolo Dalmasso, Rafael Izbicki, and Ann Lee. "Confidence sets and hypothesis testing in a likelihood-free inference setting". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 2323–2334.

[54] Constantinos Daskalakis, Gautam Chetan Kamath, and John Wright. "Which distribution distances are sublinearly testable?" In: *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 2018, pp. 2747–2764.

[55] Nicolaas Govert De Bruijn. *Asymptotic methods in analysis*. Vol. 4. Courier Corporation, 1981.

[56] Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. "Max-sliced wasserstein distance and its use for gans". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10648–10656.

[57] Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. "Generative modeling using the sliced wasserstein distance". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3483–3491.

[58] Luc Devroye, László Györfi, and Gábor Lugosi. "A note on robust hypothesis testing". In: *IEEE Transactions on Information Theory* 48.7 (2002), pp. 2111–2114.

[59] Luc Devroye and Gábor Lugosi. *Combinatorial Methods in Density Estimation*. Springer Series in Statistics. Springer New York, 2012. ISBN: 9781461301257. URL: https://books.google.com/books?id=lQEMCAAAQBAJ.

[60] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2001.

[61] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. "The total variation distance between high-dimensional Gaussians". In: *arXiv preprint arXiv:1810.08693* 6 (2018).

[62] Ilias Diakonikolas, Themis Gouleakis, Daniel M Kane, John Peebles, and Eric Price. "Optimal testing of discrete distributions with high probability". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 2021, pp. 542–555.

[63] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. "Sample-optimal identity testing with high probability". In: *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2018.

[64] Ilias Diakonikolas and Daniel M Kane. "A new approach for testing properties of discrete distributions". In: *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2016, pp. 685–694.

[65] Thomas G Dietterich and Ghulum Bakiri. "Solving multiclass learning problems via error-correcting output codes". In: *Journal of artificial intelligence research* 2 (1994), pp. 263–286.

[66] Peter J Diggle and Richard J Gratton. "Monte Carlo methods of inference for implicit statistical models". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 46.2 (1984), pp. 193–212.

[67] Conor Durkan, Iain Murray, and George Papamakarios. "On contrastive learning for likelihood-free inference". In: *International conference on machine learning*. PMLR. 2020, pp. 2771–2781.

[68] Michael Sergeevich Ermakov. "Minimax detection of a signal in a Gaussian white noise". In: *Theory of Probability & Its Applications* 35.4 (1991), pp. 667–679.

[69] Sergio Escalera, David MJ Tax, Oriol Pujol, Petia Radeva, and Robert PW Duin. "Subclass problem-dependent design for error-correcting output codes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.6 (2008), pp. 1041–1054.

[70] Alexander Fengler, Lakshmi N Govindarajan, Tony Chen, and Michael J Frank. "Likelihood approximation networks (LANs) for fast inference of simulation models in cognitive neuroscience". In: *Elife* 10 (2021), e65074.

[71] Andrey Feuerverger. "A consistent test for bivariate dependence". In: *International Statistical Review/Revue Internationale de Statistique* (1993), pp. 419–433.

[72] Jerome Friedman. *On multivariate goodness-of-fit and two-sample testing*. Tech. rep. Citeseer, 2004.

[73] Stefano Frixione, Paolo Nason, and Carlo Oleari. "Matching NLO QCD computations with parton shower simulations: the POWHEG method". In: *Journal of High Energy Physics* 2007.11 (2007), p. 070.

[74] Patrik Róbert Gerber, Yanjun Han, and Yury Polyanskiy. *Minimax optimal testing by classification*. PMLR, 2023. arXiv: 2306.11085 [math.ST].

[75] Patrik Róbert Gerber, Tianze Jiang, Yury Polyanskiy, and Rui Sun. "Density estimation using the perceptron". In: *arXiv preprint arXiv:2312.17701* (2023).

[76] Patrik Róbert Gerber, Tianze Jiang, Yury Polyanskiy, and Rui Sun. "Kernel-Based Tests for Likelihood-Free Hypothesis Testing". In: *Advances in Neural Information Processing Systems* 36 (2024).

[77] Patrik Róbert Gerber and Yury Polyanskiy. "Likelihood-free hypothesis testing". In: *arXiv preprint arXiv:2211.01126* (2022).

[78] Patrik Róbert Gerber and Yury Polyanskiy. "Likelihood-free hypothesis testing". In: *CoRR* abs/2211.01126 (2022). DOI: 10.48550/arXiv.2211.01126. arXiv: 2211.01126. URL: https://doi.org/10.48550/arXiv.2211.01126.

[79] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

[80] Oded Goldreich. *Introduction to property testing*. Cambridge University Press, 2017.

[81] Oded Goldreich. "On testing expansion in bounded-degree graphs". In: *Electronic Colloquium on Computational Complexity (ECCC)*. Vol. 20. 2000.

[82] Oded Goldreich. "The uniform distribution is complete with respect to testing identity to a fixed distribution." In: *Electron. Colloquium Comput. Complex.* Vol. 23. 2016, p. 15.

[83] Polina Golland and Bruce Fischl. "Permutation tests for classification: towards statistical significance in image-based studies". In: *Biennial international conference on information processing in medical imaging*. Springer. 2003, pp. 330–341.

[84] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[85] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. "A kernel two-sample test". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.

[86] Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. "A fast, consistent kernel two-sample test". In: *Advances in neural information processing systems* 22 (2009).

[87] Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. "Optimal kernel choice for large-scale two-sample tests". In: *Advances in neural information processing systems* 25 (2012).

[88] Shivam Gupta and Eric Price. "Sharp constants in uniformity testing via the huber statistic". In: *Conference on Learning Theory*. PMLR. 2022, pp. 3113–3192.

[89] Venkatesan Guruswami and Prasad Raghavendra. "Hardness of learning halfspaces with noise". In: *SIAM Journal on Computing* 39.2 (2009), pp. 742–765.

[90] Michael Gutman. "Asymptotically optimal classification for multiple tests with empirically observed statistics". In: *IEEE Transactions on Information Theory* 35.2 (1989), pp. 401–408.

[91] Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. "Likelihood-free inference via classification". In: *Statistics and Computing* 28.2 (2018), pp. 411–425.

[92] Mahdi Haghifam, Vincent YF Tan, and Ashish Khisti. "Sequential classification with empirically observed statistics". In: *IEEE Transactions on Information Theory* 67.5 (2021), pp. 3095–3113.

[93] Haiyun He, Lin Zhou, and Vincent YF Tan. "Distributed detection with empirically observed statistics". In: *IEEE Transactions on Information Theory* 66.7 (2020), pp. 4349–4367.

[94] Simon Hediger, Loris Michel, and Jeffrey Näf. "On the use of random forest for two-sample testing". In: *Computational Statistics & Data Analysis* 170 (2022), p. 107435.

[95] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[96] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. "Kernel methods in machine learning". In: (2008).

[97] Philip B Holden, Neil R Edwards, James Hensman, and Richard D Wilkinson. "ABC for climate: dealing with expensive simulators". In: *Handbook of approximate Bayesian computation* (2018), pp. 569–95.

[98] Hung-Wei Hsu and I-Hsiang Wang. "On binary statistical classification from mismatched empirically observed statistics". In: *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2020, pp. 2533–2538.

[99] Cheng Huang and Xiaoming Huo. "An efficient and distribution-free two-sample test based on energy statistics and random projections". In: *arXiv preprint arXiv:1707.04602* (2017).

[100] Dayu Huang. "Hypothesis Testing and Learning with Small Samples". PhD thesis. Citeseer, 2012.

[101] Dayu Huang. "Hypothesis testing and learning with small samples". PhD thesis. University of Illinois at Urbana-Champaign, 2013.

[102] Dayu Huang and Sean Meyn. "Classification with high-dimensional sparse samples". In: *2012 IEEE International Symposium on Information Theory Proceedings*. IEEE. 2012, pp. 2586–2590.

[103] Dayu Huang and Sean Meyn. "Generalized error exponents for small sample universal hypothesis testing". In: *IEEE transactions on information theory* 59.12 (2013), pp. 8157–8181.

[104] Peter J Huber. "A robust version of the probability ratio test". In: *The Annals of Mathematical Statistics* (1965), pp. 1753–1758.

[105] Peter J Huber and Volker Strassen. "Minimax tests and the Neyman-Pearson lemma for capacities". In: *The Annals of Statistics* (1973), pp. 251–263.

[106] Il'dar A Ibragimov and Rafail Z Khas'minskii. "Asymptotic properties of some nonparametric estimates in a Gaussian white noise". In: *Proc. 3rd Summer School on Probab. Theory and Math. Stat. Varna 1978* (1980), pp. 31–64.

[107] Il'dar A Ibragimov and Rafail Z Khas'minskii. "Estimation of distribution density". In: *Journal of Soviet Mathematics* 21 (1983), pp. 40–57.

[108] Il'dar A Ibragimov and Rafail Z Khas'minskii. "On the estimation of an infinite-dimensional parameter in Gaussian white noise". In: *Doklady Akademii Nauk*. Vol. 236. 5. Russian Academy of Sciences. 1977, pp. 1053–1055.

[109] Il'dar A Ibragimov and Rafail Z Khas'minskii. *Statistical estimation: asymptotic theory*. Springer Science & Business Media, 1981.

[110] Yuri I Ingster. "Minimax testing of nonparametric hypotheses on a distribution density in the L_p metrics". In: *Theory of Probability & Its Applications* 31.2 (1987), pp. 333–337.

[111] Yuri I Ingster. "On minimax distinguishability of families of nonparametric hypotheses". In: *Dokl. Akad. Nauk SSSR* 267 (3 1982), pp. 536–539.

[112] Yuri I Ingster. "On the minimax nonparametric detection of signals in white gaussian noise". In: *Problemy Peredachi Informatsii* 18.2 (1982), pp. 61–73.

[113] Yuri I Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Vol. 169. Springer Science & Business Media, 2003.

[114] Rafael Izbicki, Ann Lee, and Chad Schafer. "High-dimensional density ratio estimation with extensions to approximate likelihood computation". In: *Artificial intelligence and statistics*. PMLR. 2014, pp. 420–429.

[115] Philippe Jacquet and Wojciech Szpankowski. "Analytical depoissonization and its applications". In: *Theoretical Computer Science* 201.1-2 (1998), pp. 1–62.

[116] Zeyu Jia, Yury Polyanskiy, and Yihong Wu. "Entropic characterization of optimal rates for learning Gaussian mixtures". In: *arXiv preprint arXiv:2306.12308* (2023).

[117] Bai Jiang, Tung-yu Wu, Charles Zheng, and Wing H Wong. "Learning summary statistic for approximate Bayesian computation via deep neural network". In: *Statistica Sinica* (2017), pp. 1595–1618.

[118] Wittawat Jitkrittum, Heishiro Kanagawa, Patsorn Sangkloy, James Hays, Bernhard Schölkopf, and Arthur Gretton. "Informative features for model comparison". In: *Advances in Neural Information Processing Systems* 31 (2018).

[119] Iain M. Johnstone. *Gaussian estimation: Sequence and wavelet models*. 2019.

[120] Jørn Justesen. "Class of constructive asymptotically good algebraic codes". In: *IEEE Transactions on information theory* 18.5 (1972), pp. 652–656.

[121] Rob Kaas and Jan M Buhrman. "Mean, median and mode in binomial distributions". In: *Statistica Neerlandica* 34.1 (1980), pp. 13–18.

[122] Gautam Chetan Kamath. "Modern challenges in distribution testing". PhD thesis. Massachusetts Institute of Technology, 2018.

[123] Benjamin G Kelly, Thitidej Tularak, Aaron B Wagner, and Pramod Viswanath. "Universal hypothesis testing in the learning-limited regime". In: *2010 IEEE International Symposium on Information Theory*. IEEE. 2010, pp. 1478–1482.

[124] Benjamin G Kelly, Aaron B Wagner, Thitidej Tularak, and Pramod Viswanath. "Classification of homogeneous data with large alphabets". In: *IEEE transactions on information theory* 59.2 (2012), pp. 782–795.

[125] Arlene KH Kim and Adityanand Guntuboyina. "Minimax bounds for estimating multivariate Gaussian location mixtures". In: *Electronic Journal of Statistics* 16.1 (2022), pp. 1461–1484.

[126] Ilmun Kim, Aaditya Ramdas, Aarti Singh, and Larry Wasserman. "Classification accuracy as a proxy for two-sample testing". In: *The Annals of Statistics* 49.1 (2021), pp. 411–434.

[127] Diederik P Kingma and Max Welling. "An introduction to variational autoencoders". In: *Foundations and Trends® in Machine Learning* 12.4 (2019), pp. 307–392.

[128] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1412.6980.

[129] Szymon Knop, Marcin Mazur, Jacek Tabor, Igor Podolak, and Przemys*ℓawSpurek*. "Sliced generative models". In: *arXiv preprint arXiv:1901.10417* (2019).

[130] Szymon Knop, Jacek Tabor, Igor Podolak, Marcin Mazur, et al. "Cramer-Wold auto-encoder". In: *Journal of Machine Learning Research* 21.164 (2020), pp. 1–28.

[131] Soheil Kolouri, Kimia Nadjahi, Shahin Shahrampour, and Umut Şimşekli. "Generalized sliced probability metrics". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 4513–4517.

[132] Soheil Kolouri, Gustavo K Rohde, and Heiko Hoffmann. "Sliced wasserstein distance for learning gaussian mixture models". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3427–3436.

[133] Soheil Kolouri, Yang Zou, and Gustavo K Rohde. "Sliced Wasserstein kernels for probability distributions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 5258–5267.

[134] Eun Bae Kong and Thomas G Dietterich. "Error-correcting output coding corrects bias and variance". In: *Machine learning proceedings 1995*. Elsevier, 1995, pp. 313–321.

[135] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images". In: *University of Toronto* (2009).

[136] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. "The CIFAR-10 dataset (2014)". In: *Online: http://www. cs. toronto. edu/kriz/cifar. html* 55 (2020).

[137] Joseph Lam-Weil, Alexandra Carpentier, and Bharath K Sriperumbudur. "Local minimax rates for closeness testing of discrete distributions". In: *Bernoulli* 28.2 (2022), pp. 1179–1197.

[138] Naum S. Landkof. *Foundations of modern potential theory*. Vol. 180. Springer, 1972.

[139] Tong Li and Ming Yuan. "On the optimality of Gaussian kernel based nonparametric tests against smooth alternatives". In: *arXiv preprint arXiv:1909.03302* (2019).

[140] Yi Li. *Goodness-of-fit tests for Dirichlet distributions with applications*. Bowling Green State University, 2015.

[141] Tengyuan Liang. "How well generative adversarial networks learn distributions". In: *The Journal of Machine Learning Research* 22.1 (2021), pp. 10366–10406.

[142] Luca Lista. *Statistical methods for data analysis in particle physics*. Vol. 941. Springer, 2017.

[143] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J Sutherland. "Learning deep kernels for non-parametric two-sample tests". In: *International conference on machine learning*. PMLR. 2020, pp. 6316–6326.

[144] David Lopez-Paz and Maxime Oquab. "Revisiting Classifier Two-Sample Tests". In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=SJkXfE5xx.

[145] Subhash Chandra Malik and Savita Arora. *Mathematical analysis*. New Age International, 1992.

[146] Enno Mammen and Alexandre B Tsybakov. "Smooth discrimination analysis". In: *The Annals of Statistics* 27.6 (1999), pp. 1808–1829.

[147] Youssef Marzouk, Zhi Ren, Sven Wang, and Jakob Zech. "Distribution learning via neural differential equations: a nonparametric statistical perspective". In: *arXiv preprint arXiv:2309.01043* (2023).

[148] Henryk Minc and Leroy Sathre. "Some inequalities involving $(r!)^{1/r}$". In: *Proceedings of the Edinburgh Mathematical Society* 14.1 (1964), pp. 41–46.

[149] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

[150] Tamás F Móri, Gábor J Székely, and Maria L Rizzo. "On energy tests of normality". In: *Journal of Statistical Planning and Inference* 213 (2021), pp. 1–15.

[151] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. "Kernel mean embedding of distributions: A review and beyond". In: *Foundations and Trends® in Machine Learning* 10.1-2 (2017), pp. 1–141.

[152] Kimia Nadjahi, Alain Durmus, Lénaıc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. "Statistical and topological properties of sliced probability divergences". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 20802–20812.

[153] Francis J Narcowich and Joseph D Ward. "Norm estimates for the inverses of a general class of scattered-data radial-function interpolation matrices". In: *Journal of Approximation Theory* 69.1 (1992), pp. 84–109.

[154] Bobak Nazer, Or Ordentlich, and Yury Polyanskiy. "Information-distilling quantizers". In: *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2017, pp. 96–100.

[155] Peter Neal. "Efficient likelihood-free Bayesian computation for household epidemics". In: *Statistics and Computing* 22 (2012), pp. 1239–1256.

[156] DJ Newman and HS Shapiro. "Jackson's theorem in higher dimensions". In: *On Approximation Theory/Über Approximationstheorie*. Springer, 1964, pp. 208–219.

[157] Jerzy Neyman and Egon Sharpe Pearson. "IX. On the problem of the most efficient tests of statistical hypotheses". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231.694-706 (1933), pp. 289–337.

[158] Sloan Nietert, Ziv Goldfeld, Ritwik Sadhu, and Kengo Kato. "Statistical, robustness, and computational guarantees for sliced wasserstein distances". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 28179–28193.

[159] Patrick Ofosuhene. "The energy goodness-of-fit test for the inverse Gaussian distribution". PhD thesis. Bowling Green State University, 2020.

[160] Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. "Diffusion models are minimax optimal distribution estimators". In: *arXiv preprint arXiv:2303.01861* (2023).

[161] Seunghoon Paik, Michael Celentano, Alden Green, and Ryan J Tibshirani. "Maximum Mean Discrepancy Meets Neural Networks: The Radon-Kolmogorov-Smirnov Test". In: *arXiv preprint arXiv:2309.02422* (2023).

[162] Liam Paninski. "A coincidence-based test for uniformity given very sparsely sampled discrete data". In: *IEEE Transactions on Information Theory* 54.10 (2008), pp. 4750–4755.

[163] George Papamakarios, David Sterratt, and Iain Murray. "Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 837–848.

[164] Ankit Pensia, Varun Jog, and Po-Ling Loh. "Communication-constrained hypothesis testing: Optimality, robustness, and reverse data processing inequalities". In: *arXiv preprint arXiv:2206.02765* (2022).

[165] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. "Machine learning classifiers and fMRI: a tutorial overview". In: *Neuroimage* 45.1 (2009), S199–S209.

[166] Iosif Pinelis. "Positive-part moments via characteristic functions, and more general expressions". In: *Journal of Theoretical Probability* 31 (2018), pp. 527–555.

[167] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. *Diffusers: State-of-the-art diffusion models*. https://github.com/huggingface/diffusers. 2022.

[168] Y Polyanskiy and Y Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2023+.

[169] Yury Polyanskiy and Zeyu Jia. Personal communication. Feb. 2024.

[170] Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. "Feature learning in neural networks and kernel machines that recursively learn features". In: *arXiv preprint arXiv:2212.13881* (2022).

[171] Aaditya Ramdas, Nicolás Garcıa Trillos, and Marco Cuturi. "On wasserstein two-sample testing and related families of nonparametric tests". In: *Entropy* 19.2 (2017), p. 47.

[172] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125* ().

[173] Frigyes Riesz and Béla Szöke Nagy. *Functional analysis*. Courier Corporation, 2012.

[174] Maria L Rizzo and John T Haman. "Expected distances and goodness-of-fit for the asymmetric Laplace distribution". In: *Statistics & Probability Letters* 117 (2016), pp. 158–164.

[175] Pau Rodrıguez, Miguel A Bautista, Jordi Gonzalez, and Sergio Escalera. "Beyond one-hot encoding: Lower dimensional target embedding". In: *Image and Vision Computing* 75 (2018), pp. 21–31.

[176] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.

[177] Bero Roos. "On the rate of multivariate Poisson convergence". In: *Journal of Multivariate Analysis* 69.1 (1999), pp. 120–134.

[178] Harold Sackrowitz and Ester Samuel-Cahn. "P Values as Random Variables-Expected P Values". In: *The American Statistician* 53.4 (1999), pp. 326–331. ISSN: 00031305. URL: http://www.jstor.org/stable/2686051 (visited on 01/19/2023).

[179] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, June 2001. ISBN: 9780262256933. DOI: 10.7551/mitpress/4175.001.0001. URL: https://doi.org/10.7551/mitpress/4175.001.0001.

[180] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. "Equivalence of distance-based and RKHS-based statistics in hypothesis testing". In: *The annals of statistics* (2013), pp. 2263–2291.

[181] Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabás Póczos. "Nonparametric density estimation under adversarial losses". In: *Advances in Neural Information Processing Systems* 31 (2018).

[182] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian computation*. CRC Press, 2018.

[183] Torbjörn Sjöstrand, Stephen Mrenna, and Peter Skands. "PYTHIA 6.4 physics and manual". In: *Journal of High Energy Physics* 2006.05 (2006), p. 026.

[184] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. "Score-based generative modeling through stochastic differential equations". In: *arXiv preprint arXiv:2011.13456* (2020).

[185] Elias M Stein. *Singular integrals and differentiability properties of functions*. Princeton university press, 1970.

[186] Elias M Stein and Guido Weiss. *Introduction to Fourier analysis on Euclidean spaces*. Vol. 1. Princeton university press, 1971.

[187] Danica J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alexander J. Smola, and Arthur Gretton. "Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL: https://openreview.net/forum?id=HJWHIKqgl.

[188] Gábor J Székely and Maria L Rizzo. "A new test for multivariate normality". In: *Journal of Multivariate Analysis* 93.1 (2005), pp. 58–80.

[189] Gábor J Székely and Maria L Rizzo. "Brownian distance covariance". In: *The annals of applied statistics* (2009), pp. 1236–1265.

[190] Gábor J Székely and Maria L Rizzo. "Energy statistics: A class of statistics based on distances". In: *Journal of statistical planning and inference* 143.8 (2013), pp. 1249–1272.

[191] Gábor J Székely and Maria L Rizzo. "Partial distance correlation with methods for dissimilarities". In: (2014).

[192] Gábor J Székely, Maria L Rizzo, et al. "Testing for equal distributions in high dimension". In: *InterStat* 5.16.10 (2004), pp. 1249–1272.

[193] Gábor J Székely and Maria L Rizzo. *The energy of data and distance correlation*. CRC Press, 2023.

[194] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. "Measuring and testing dependence by correlation of distances". In: (2007).

[195] GáborJ Székely. "Potential and kinetic energy in statistics". In: *Lecture Notes, Budapest Institute* (1989).

[196] Owen Thomas, Ritabrata Dutta, Jukka Corander, Samuel Kaski, and Michael U Gutmann. "Likelihood-free inference by ratio estimation". In: *Bayesian Analysis* 17.1 (2022), pp. 1–31.

[197] Stefan Tiegel. "Hardness of Agnostically Learning Halfspaces from Worst-Case Lattice Problems". In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Ed. by Gergely Neu and Lorenzo Rosasco. Vol. 195. Proceedings of Machine Learning Research. PMLR, 2023, pp. 3029–3064. URL: https://proceedings.mlr.press/v195/tiegel23a.html.

[198] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. 1st. Springer Publishing Company, Incorporated, 2008. ISBN: 0387790519.

[199] Gregory Valiant and Paul Valiant. "An automatic inequality prover and instance optimal identity testing". In: *SIAM Journal on Computing* 46.1 (2017), pp. 429–455.

[200] Paul Valiant. "Testing symmetric properties of distributions". In: *SIAM Journal on Computing* 40.6 (2011), pp. 1927–1968.

[201] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

[202] Sven Wang and Youssef Marzouk. "On minimax density estimation via measure transport". In: *arXiv preprint arXiv:2207.10231* (2022).

[203] G.N. Watson. *A Treatise on the Theory of Bessel Functions*. Cambridge Mathematical Library. Cambridge University Press, 1995. ISBN: 9780521483919.

[204] Shuo Yang, Ping Luo, Chen Change Loy, Kenneth W Shum, and Xiaoou Tang. "Deep representation learning with target coding". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1. 2015.

[205] Yuhong Yang. "Minimax nonparametric classification. I. Rates of convergence". In: *IEEE Transactions on Information Theory* 45.7 (1999), pp. 2271–2284.

[206] Yannis G Yatracos. "Rates of convergence of minimum distance estimators and Kolmogorov's entropy". In: *The Annals of Statistics* 13.2 (1985), pp. 768–774.

[207] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. "Advances in variational inference". In: *IEEE transactions on pattern analysis and machine intelligence* 41.8 (2018), pp. 2008–2026.

[208] Ruichong Zhang. "Cramer Type Distances for Learning Gaussian Mixture Models by Gradient Descent". In: *arXiv preprint arXiv:2307.06753* (2023).

[209] Lin Zhou, Vincent YF Tan, and Mehul Motani. "Second-order asymptotically optimal statistical classification". In: *Information and Inference: A Journal of the IMA* 9.1 (2020), pp. 81–111.

[210] Chao-Zhe Zhu, Yu-Feng Zang, Qing-Jiu Cao, Chao-Gan Yan, Yong He, Tian-Zi Jiang, Man-Qiu Sui, and Yu-Feng Wang. "Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder". In: *Neuroimage* 40.1 (2008), pp. 110–120.

[211] Abram A Zinger, Ashot V Kakosyan, and Lev B Klebanov. "A characterization of distributions by mean values of statistics and certain probabilistic metrics". In: *Journal of Soviet Mathematics* 59.4 (1992), pp. 914–920.

[212] Jacob Ziv. "On classification with empirically observed statistics and universal data compression". In: *IEEE Transactions on Information Theory* 34.2 (1988), pp. 278–286.